1994 BENCHMARK TESTS FOR THE ARPA SPOKEN LANGUAGE PROGRAM

David S. Pallett, Jonathan G. Fiscus, William M. Fisher,
John S. Garofolo, Bruce A. Lund, Alvin Martin, and Mark A. Przybocki

National Institute of Standards and Technology (NIST)
Room A216, Building 225 (Technology)
Gaithersburg, MD 20899

ABSTRACT

This paper reports results obtained in benchmark tests conducted within the Spoken Language portion of the ARPA Human Language Technology Program in November and December of 1994. As in previous years [1], in addition to ARPA contractors, participants included a number of "volunteers", including foreign participants from Canada, France, Germany, and the United Kingdom. The body of the paper includes an outline of the structure of the tests and presents highlights and discussion of selected results. Detailed tabulations of reported "official" results, and additional explanatory text appear in the Appendix.

1. INTRODUCTION

Benchmark tests were implemented within the ARPA Human Language Technology research program during the period November - December 1994. In this year's large-vocabulary Continuous Speech Recognition (CSR) technology tests, the scope was broadened to include not only the use of Wall Street Journal-based Continuous Speech Recognition (WSJ-CSR) Corpus texts, but also other texts involving North American Business (NAB) news. "Read" speech and language models were derived from this new source material.

This year the vocabulary required for complete coverage of some of the Hub test material was unlimited in the sense that a substantial fraction (126/316) of the test utterances included at least one Out-Of-Vocabulary (OOV) word not included within the coverage of the "official" 20K-word trigram language model. New words, or OOVs, have their origin in NAB news items occurring after a designated cutoff date for the generation of lexicons and language models.

The training and test material, for the CSR tests, as in recent years, was collected at SRI International (SRI) under contract to the Linguistic Data Consortium (LDC). Spoken language understanding technology tests made use of ARPA Air Travel Information System (ATIS) material collected at several sites, processed at NIST, annotated at SRI, and provided to participating members of the LDC.

2. CONTINUOUS SPEECH RECOGNITION (CSR) TESTS

2.1. 1994 CSR Test Paradigm

2.1.1 The Hubs

All sites participating in the CSR tests were required to submit results for at least one "Hub" or baseline test. The Hub 1 (H1) tests were intended to measure basic speaker-independent performance and required use of read NAB News material, collected with a standard headset-boom-mounted, close-talking microphone (i.e., Sennheiser HMD-410-6).

This year's H1 test material included some out-of-vocabulary (OOV) words: words that did not occur within the lexicon defined by an agreed-upon standard 20K trigram language model. The inclusion of OOV material was intended to yield insights into the behavior of systems presented with new words.

Also new this year was a second Hub (H2) test, to provide a baseline for the recognition of unlimited-vocabulary NAB News read speech over long-distance telephone lines, using digital "T1" service (8-bit mu law) at the data collection end, and unconstrained handset/microphone, electronics and background noise environment at the subject end. The use of any grammar or acoustic training predating mid-June, 1994 was permitted.

2.1.2 The Spokes

As in the 1993 tests, the Spoke tests were intended to support a number of different challenges involving adaptation and compensation. For many of the spokes, a 5K language model was used. Some of the spokes defined in last year's tests had no participants this year; however, the numbering used last year was unchanged. Thus no results are reported, this year, for last year's Spokes 6,7, and 8.

Spokes 2, 3 and 4 supported approaches to problems in various types of adaptation: Spoke 2 -- adaptation to new news topics not found in the training material, Spoke 3 -- adaptation for "recognition outliers" (i.e., non-native speakers), and Spoke 4 -- incremental speaker adaptation.

Spokes 5 and 10 supported problems in noise and channel compensation: Spoke 5 -- unsupervised channel compensation (using a variety of microphones), and Spoke 10, a new ARPA spoke this year -- compensation on data "corrupted" with additive noise.

Spoke 9, as in last year's tests, dealt with spontaneous dictation of business news stories.

Additional details of the design of the CSR Benchmark tests appear in Kubala [2].

2.2. WSJ-CSR Summary Highlights of Results

Hub 1: There were 15 research "teams" from 14 sites participating in the Hub 1 tests.

In the Hub 1 baseline (contrastive) condition (H1C1), word error rates ranged from 10.5% to 22.8%, and for the primary condition (H0P0), in which the use of any grammar or acoustic training predating mid-June, 1994 was permitted, error rates ranged from 7.2% to 22.8%.

The Hub 1 C1 baseline condition involved use of a "standard" trigram language model and restriction of the acoustic training data to either of two equal-sized datasets [2]. The lowest word error rate in this baseline condition (10.5%) was achieved by the Cambridge University Engineering Department's (CU) HMM Toolkit developers, using a "continuous mixture density, tied state, gender dependent cross-word context dependent" system.

This group also achieved the lowest word error rate (7.2%) in the Hub 1 Primary condition (H1P0). This condition permitted the use of any acoustic training data and any language model. The Cambridge H1P0 system also incorporated incremental unsupervised adaptation and a 4-gram language model with a 65K-wordlist and pronunciations generated by a text-to-speech system mapped to use the LIMSI phone set. The "P0:C1" reduction in word error rate (from 10.5% to 7.2%) amounts to 31.6% reduction, and was shown to be statistically significant using several significance tests.

Hub 2: Three sites (CMU, LIMSI and SRI International) participated in the Hub 2 tests involving telephone NAB News, with error rates ranging between 22.5% and 24.6%. Performance differences between systems were not, in general, significant. According to SRI, the SRI system, with the lowest word error rate, "is very similar to SRI's system used for the Hub 1-C1 evaluation", but differences include the use of telephone acoustic training from the Linguistic Data Consortium's Macrophone and Switchboard corpora and "more efficient gender selection".

Spoke 2: Spoke 2 involved adaptation to new news topics not found in the training material. CMU was the only site participating in this spoke. The (incomplete) results reported to NIST as of January 17, 1995 do not include the use of topic-specific training, and are thus largely inconclusive.

Spoke 3: Spoke 3 involved the use of unsupervised adaptation for "read" speech from non-native speakers. BBN, CU, and SRI participated in this spoke. With adaptation disabled, word error rates ranged from 20.7% to 26.1%, but with adaptation enabled, using the standard 40 sentence "rapid enrollment data", word error rates were typically halved by each of the sites.

Spoke 4: Spoke 4 involved incremental speaker adaptation, using test material consisting of 100-utterance sets from 4 different speakers. BBN, CU and Dragon Systems participated in this spoke. Enabling unsupervised incremental adaptation (P0) provided varying degrees of reduction in error rate when compared to the corresponding "adaptation disabled" condition (C1) -- in some cases between 10% and 30% reduction in error rates. Additional contrastive tests involving the use of supervised incremental adaptation indicated modest additional improvements that were in almost every case not statistically significant.

Spoke 5: Spoke 5 involved the use of unsupervised "channel" compensation for a number of different microphones, other than the close-talking Sennheiser microphone. CMU was the sole participant. One CMU system used for this spoke incorporated a new algorithm called "RATZ (snR-dependent gAussian-based cepstTral normaliZation)", used to compensate for the effects of recording speech through an unknown microphone. The other CMU system used "a new version of the CDCN (Codebook-Dependent Cepstral Normalization) used to compensate for the combined effects of additive noise and unknown channel distortion". The enabling of compensation (as noted in the P0:C1 contrast) achieved approximately 20% reduction in word error rate for both approaches.

Spoke 9: Spoke 9 involved spontaneously dictated business news stories, and the sole participant was BBN. A word error rate of 14.2% was achieved, using "exactly the same system as used in the H1-P0" tests. This system involves the use of "incremental unsupervised adaptation, state-clustered tied-mixture continuous densities, and clustered triphone states using decision trees, and no interpolation with weaker context models".

Spoke 10: Spoke 10 involved the use of compensation on data "corrupted" with additive noise at different speech Signal to Noise Ratios (SNRs). The additive noise for this spoke was recorded by SRI in automobiles travelling on freeways at speeds of approximately 60 miles per hour. Participating sites included Cambridge University, IBM, and SRI.

The Cambridge University HTK system for Spoke 10 implemented an approach called the "Parallel Model Combination (PMC)" technique, which "attempts to estimate a set of matched models by combining a fixed set of acoustic models trained on clean speech and a noise model dynamically based on the background noise". IBM used a similar approach. SRI's approach involved using "a combination of feature mapping and HMM adaptation techniques to modify [the] seed system". The mapping is done using "Probabilistic Optimum

Filtering (POF)", for various SNR levels.

All systems showed markedly increasing word error rates with decreasing SNR, but with enabled noise compensation, the rate of increase was lessened. For the SRI system, an error rate of 6.7% was obtained for the original speech data (without additive noise), and with compensation enabled, error rates of 8.4%, 9.8%, and 12.2% were obtained for the 22, 16 and 10 dB SNR conditions.

2.3. WSJ-CSR Discussion

2.3.1 Word Alignment, Scoring and Estimated Systematic Error

In this year's scoring, NIST made use of a phonologically-based dynamic programming procedure in matching the reference transcriptions to the system output (1-best) hypothesis files. It is well known that there is some systematic error (an underestimate of the true error rate) in scoring the output of continuous speech recognition systems because of limitations in the string matching algorithms. While the use of the phonologically-based dynamic programming procedure as a basis for scoring has been shown to be more accurate than the previous procedure, the use of "time-aligned" system output and reference transcriptions provides what is believed to be the most accurate scores. Fisher et al. provide details on this year's choice of the phonologically-based procedure for alignment and scoring [3].

From pragmatic considerations, what is of interest is the magnitude of systematic errors, or diffferences between the results obtained with the use of alternative alignment and scoring procedures.

To provide estimates of these errors, NIST obtained time-marked system outputs and reference transcriptions for the HTK system on the H1 data, and scored the system outputs using: (a) the traditional NIST alignment algorithm, (b) the new NIST phonological alignment algorithm, and (c) the time-marked files. The resulting estimates of word error rates were; (a) using traditional string alignment and scoring, 11.2%, (b) using phonological alignment, 11.3% (note that this differs from the number tabulated in Table A1 because these results were computed prior to adjudication), and (c) using time marks, 11.6%. The estimated magnitude of the systematic error associated with these CSR benchmark tests, using the phonological alignment algorithm, is approximately 1% to 2% of the reported word error rates, or an underestimate of the magnitude of the true error of approximately 0.1%. and 0.3%.

Proportionately larger systematic errors are associated with higher error rates.

2.3.2 Uncertainty and Significance Tests

It is important to recognize that there is considerable variability in the performance of each system for different speakers in each test subset. Simple means of error rates over a test population tell only a portion of the truth, and measures of variance are also important. The International Committee for Weights and Measures (CIPM) now recommends the use of the term, "standard uncertainty", which is equal to the positive square root of the estimated variance.[8]

In the H1P0 test, for the system with the lowest word error rate, the Cambridge University "htk1" system, the mean word error rate is 7.2%. NIST's scoring software notes, however, an associated standard uncertainty of 4.9%, and the range of error rates over the population of 20 test speakers is from 2.8% to 20.8%. Similar measures of uncertainty apply for other systems, and should be taken into consideration.

For several years, NIST has implemented a number of statistical significance tests for use in comparing the performance of different algorithms or systems. Application of these paired-comparison statistical significance tests indicate, for example, that in most, but not all, cases the performance differences between Cambridge University's HTK-based system and other H1C1 systems were statistically significant (see Table A2), with lower error rates for the htk1 system.

The most frequent exception was in the case of application of the McNemar test (which is a test on sentence error rate), for which many of the test results indicated that the differences were not significant. Because of the length of the sentence texts in the H1 test material (an average of approximately 26 words), utterance error rates were high for all systems and the performance differences between systems, measured in terms of utterance error rate, were in many cases not significantly different.

Other exceptions involved comparisons of the results for the htk1 system with those for other well-performing systems (i.e., ibm1 and sri1), and in these cases, the statistical significance tests do not reveal significant differences in performance.

2.3.3 Variability in Error Rate across Speakers and Systems

For most of the systems, two speakers (speakers 4t6 and 4td) in the H1 test set were flagged by NIST's scoring software as having (unusually high) error rates outside the expanded uncertainty range (which is the range +/- 2 standard deviations about the mean). In previous tests high error rates have often been associated with extremes of rate-of-speech, and/or "careless" speech [1,4]. Figure 1 illustrates the variability in word error rate for the 20 speakers in the H1 test set for the H1C1 condition, as a function of speaking rate. "Speaking Rate" (here inferred from simple counts of the number of words read and file duration) ranges from about 125 to 190 words/minute. The high error rates associated with speakers 4t6 and 4td are shown to be associated with fast speech.

Figure 2 indicates the variability in error rates for each of the 15 systems participating in the H1C1 tests, over the 20 speakers in test set. Each data point represents the word error rate for one of the speakers in the test set for one of the systems. Error rates for the individual speakers range from a low of 2.8% (for speaker 4t3 for the LIMSI system) to 49.6% (for speaker 4td for the KU system).

NIST's scoring software computes, for each test set, the total word error rate for all test material in a given test set -- the number of word errors divided by the number of words in the reference transcriptions -- rather than the unweighted mean word error rate for the ensemble of speakers in the test set. In effect, NIST reports a weighted mean word error rate for the speakers of the test set, where the weighting is a function of the number of word tokens spoken by each speaker.

Both median and mean word error rates are indicated in Figure 2, with the median error rates over the several systems systematically somewhat lower. When the test sets include speakers with unusually high error rates, or if one subject were to provide a large fraction of the test material (or, as in this case, there are some systems with markedly higher error rates than others), the mean may not be an appropriate measure to describe the performance over the ensemble of speakers (or systems).

2.3.4 Out-Of-Vocabulary (OOV) Effects

This year's test material based on North American Business news included a significant number of sentences with at least one OOV occurrence with regard to the standard baseline 20K open trigram backoff language model. For Hub 1, there were 194 occurrences of OOVs, comprising 2.4% of the word occurrences.

For many of the OOV occurrences, especially polysyllabic ones, the systems' responses consisted of "splits" (occurrences of contiguous substitution and insertion errors). Typical examples include the polysyllabic OOV "flywheels" resulting in "fly we'll" or "flight wheel" or, "powerbooks" resulting in "our books", and "centrifuges" in "sent refuses". These occurrences suggest that one might often expect two word errors per OOV-word occurrence, especially for polysyllabic OOVs. Of course, only a fraction of the OOV's are polysyllabic, and it is informative to look at the system response for OOV-word occurrence in more detail.

NIST partitioned the results into two subsets, one containing only "In-Vocabulary" (IV) words, and another also containing OOVs. The OOV-containing subset contained a total of 3657 word occurrences, of which the 194 OOV items comprised 5.3% of the word occurrences, in 126 utterances. The IV-containing subset contained 4526 word occurrences and 190 utterances.

We then scored the individual subsets of the H1 submissions. Table 1 shows the word error rates found for the IV- and OOV-containing subsets for both H1C1 and H1P0 tests. Note, for example, that for the cu-htk1 system, the error rate for the IV-containing subset of H1C1 was 6.4%, in contrast to 15.6% for the OOV-containing subset. It is clear that the OOV occurrence results in substantial degradation of performance.

The occurrence of an OOV item, and the consequent errors, will cause other errors because of the use of language models (in this case a trigram). It is of interest to estimate the number of errors that might be expected for each occurrence of an OOV item (errors/OOV).

We can assume that each of the 194 OOV occurrences led to at least one word error. To estimate the number of errors resulting for the (3659 - 194 = 3465) IV occurrences, we used the error rate applicable for the IV subset (6.4%), to estimate that 222 errors were IV-word-induced. An estimate of the number of errors for the OOV-containing subset is thus (194 + 222 = 416).

However, we actually observed another (571 - 416 = 155) errors. If we attribute these errors to the presence of OOVs, then we can attribute a total of (194 + 155 = 349) word errors to the 194 OOV occurrences, or an average of (349/194 = 1.8) word errors per OOV word occurrence.

We repeated these estimates for all systems, and found that the number of word errors per OOV-word occurrence to be relatively stable for each system, ranging from 1.7 to 2.1 errors/OOV, as indicated in Table 2.

For other calculations, performed using pre-adjudicated reference answers, the ratio is somewhat higher, because most of the errors involving compound words were "forgiven" in the process of adjudication. Also note that using pre-adjudicated reference answers, the utterance error rate ("U.E.") in Table 1, for the OOV-containing subset of H1C1, would be 100% for all systems, as expected.

There are, of course, alternative ways to estimate this ratio. Kubala notes that a simple way, that attributes all errors in the proximity of an OOV-occurrence to the OOV, is based on counting "each word error occurring within a contiguous sequence of word errors containing at least one OOV" [7]. For the H1C1 test and the bbn1 system, Kubala found an average of 2.1 errors/OOV, and for the BBN H1P0 system, with a larger lexicon, he cites 152 word errors aligned to 69 OOV word occurrences, or 2.2 errors/OOV.

Still another way to study this phenomena is to look in detail at the aligned strings, and to estimate the local error rate in the vicinity of an OOV. NIST made use of time-aligned data provided by the HTK developers to investigate this phenomenon, looking at error rates in the 5-word region surrounding OOV word occurrences (strings containing two IV words on both sides of an OOV), and comparing the results for these strings with other strings not containing any OOVs.

Table 3 shows the results of our study. For strings not containing OOV items, note that the percentage of words correctly recognized, in this case, is relatively stable throughout the strings (92.9%, 93.0%, 93.3%, 93.3%, and 93.1%). For strings containing OOVs, however, note that the percentage of words correctly recognized drops markedly in the vicinity of the OOV (90.1%, 72.8%, 0.0% ((at the OOV occurrence)), 74.2%, and 85.4%). Note the evidence of degraded performance even for one word before and two words after the OOV occurrence.

Having noted that word error rates range from 15.6% to 25.4% for the OOV-containing subset of H1CI data, and recognizing that several sites built systems with lexicons larger than 20K for the H1P0 test (as large as 65K for the cu-htk1 system), it is of interest to look at error rates for the IV and OOV subsets for the P0 systems.

Table 1 shows, for example, that the performance of the P0 systems, in general, improved when compared to the C1 systems by an amount that is, in many cases, not statistically significant for the IV subset (e.g., note the P0-IV:C1-IV comparisons and significance tests). It is noteworthy that the att1 and phil-th2 P0 systems achieved significant reductions in IV-subset word error. Note however that performance improved markedly for the P0 systems for the OOV-containing subset (e.g., note the P0-OOV:C1-OOV comparisons and significance tests), most probably because of increased lexical coverage for the PO systems. Typical reductions in word error rate for this subset are in the range of 20 to 30 percent.

2.3.5 SNR Properties and Compensation (Spokes 5 and 10)
As described earlier, Spoke 5 involved the use of unsupervised "channel" compensation or a number of different microphones, and Spoke 10 involved the use of compensation on data "corrupted" with additive noise.

For Spoke 5, there were 10 different microphones: 4 tie-clip microphones (1 of which was a wireless mike), 3 stand-mounted or gooseneck microphones, 2 "flat" desktop mikes, and 1 hand-held mike. The group at SRI who collected the data reported that placement of the tie-clip microphones was such that these microphones were 7 to 10 inches below the subject's mouth. Microphone-to-mouth distances for the stand-mounted microphones were 12 to 18 inches, and for the desktop microphones, 16 to 30 inches. The SRI group reported that the hand-held mike "tended to pickup a significant amount of hum from the terminal when placed in a stand" and was for that reason used as a handheld mike, with subjects instructed to hold the microphones in front of them, and about 5 to 8 inches from their mouth. SRI also reported that one of the inexpensive stand-mounted mikes, packaged with a commercial product, "picked up a very loud buzzing noise when placed anywhere within about 3 inches" of a terminal, and was for that reason used only as a stand-mounted microphone to the left of the keyboard.

For the Spoke 5 test set, each of the different microphones was used by only two different speakers, so that channel effects are somewhat confounded by speaker effects, as noted below.

SNR measurements performed at NIST, using NIST's SNR software (which estimates the ratio of peak-speech-power-level to mean-noise-power-level), indicate that with use of A-weighting (a standardized form of low-frequency de-emphasis frequently used in noise measurements), measured SNRs were typically 20 to 30 dB. Appreciable low frequency energy, with components in the 20 to 50 Hz range, was found in the data from one of the lapel microphones and the hand-held microphone. This energy, which might originate in either physical vibration (shaking hands) or ventilation system noise, resulted in broadband SNRs as low as 14 to 20 dB for those two microphones. For the four tie-clip microphones, A-weighted SNRs ranged from approximately 27 dB to 36 dB, with the lower SNR obtained for the microphone using the wireless transmitter and receiver.

Without enabling compensation, CMU achieved a 12.4% word error using the data collected with 10 different microphones, in contrast to 6.7% for the corresponding data from the close talking standard Sennheiser microphone. This amounts to an 85% increase in word error, if compared to the results using the close talking microphone. However, implementing either of the two CMU compensation algorithms reduces the error rate for the alternative microphone set to 9.9% and 9.7%, approximately a 20% reduction in word error. Note that this still amounts to a 46% increase in word error rate, if compared to the results using the close talking microphone, but less of an increase than without enabling compensation.

Even though the P0C1 reduction in weighted mean word error amounts to 20%, individual reductions range from -12.5% to 69%. There is also appreciable variation in the reduction for the two speakers sharing the same alternative microphone (e.g., reductions of 44% and -2.9% for two speakers using one of the desktop microphones).

Spoke 10 involved the use of compensation on data (obtained from the Sennheiser close-talking microphone) "corrupted" with additive noise, as described earlier. Use of a one-minute sample of noise for adaptation was permitted. The test participants had specified that an overall 10-speaker RMS speech level be determined, and that long-term RMS measures of the automobile interior noise be determined, and that these measures be used to determine the amount of additive noise, relative to the speech level. Consequently, depending on varying vocal effort and instantaneous noise level, the SNR within a nominal SNR condition varied over an 8 dB range over the ensemble of test speakers. The test data included three (nominal) A-weighted signal to noise ratios (SNRs): 22 dB, 16 dB, and 10 dB. RMS A-weighted measures of both speech and noise energy were used in calculating the SNRs. Broadband (unweighted) SNRs would be approximately 6 dB worse (i.e., ranging from approximately 16 dB to 4 dB).

Despite the variation of SNR over the ensemble of test speakers within each test set, relatively low error rates, as well as evidence of effective compensation, were found near both extremes of speech level (and thus SNR) within a given test subset.

As previously noted, all systems showed markedly increasing word error rates with decreasing SNR, but with enabled noise compensation, the rate of increase was lessened. There was a substantial variation in the effectiveness of the several noise compensation approaches: the three participants each achieved error rates ranging from 6.7% to 7.2% for the "clean" data, but for the worst SNR case (10 dB), error rates with compensation enabled ranged from 12.2% to 19.8%. Even the lowest word error rate for the worst SNR case (12.2%), however, represents an 82% increase in error rate relative to that for the clean speech

(6.7%).

Ebel and Picone report average word error rates of 1% for recognition and transcription of the S10 data by humans, and notes that "human performance measured in terms of word error performance did not vary significantly with SNR" [5].

3. ATIS TESTS

3.1. New Conditions

This year's ATIS tests were similar to the 1993 ATIS tests. As in prior years, tests included spontaneous speech recognition (SPREC) tests, natural language understanding (NL) tests and spoken language understanding (SLS) tests. Only unweighted NL and SLS errors are reported (i.e., incorrect answers count the same as "No Answer" responses), this year, as was the case last year, rather than citing a weighted error measure which penalized use of the "No_Answer" less heavily than a "False" answer.

3.2. Summary Test Results

For the recent ATIS tests, results were reported for systems at seven sites, including AT&T and MITRE as "volunteers".

Additional details about the test paradigm are found in another paper in this proceedings by Dahl [6]. Details about the technical approaches used by the participants, and their own analyses and comments, are to be found in other references.

SPontaneous Speech RECognition (SPREC) Tests
For the SPontaneous speech RECognition (SPREC) tests, for the subset of all answerable queries (Class A+D), the word error rates ranged from 1.9% to 14.1%. As in previous years, the lowest error rates are typically found for the subset of context-independent queries (Class A), with higher error rates for context-dependent queries (Class D), and the "unanswerable" queries (Class X).

The lowest word error rates for the subset Class A+D were achieved by two versions of the SRI DECIPHER system, which is described by SRI as "based on a progressive-search strategy, and shared Gaussian Mixture, gender-dependent Hidden Markov Models". Acoustic training for the DECIPHER systems included not only a set of approximately 20,000 spontaneous ATIS utterances, but also the Wall Street Journal "SI-284" training corpus. A second version of the DECIPHER System made use of information from the SRI ATIS GEMINI natural language system to implement rescoring of an N-best list provided by DECIPHER. Information used for rescoring is based on use of a statistical language model "based on the best analysis GEMINI can find of a recognition hypotheses as a sequence of phrases that can be mapped onto the ATIS database, skipping as few words as possible".

For the N-best system (sri4), for the subset of answerable utterances (Class A + D), for all but 2 of 36 possible paired-comparison tests, significant SPREC test differences were indicated, with lower error rates for the sri4 system.

Natural Language (NL) Tests
For the Natural Language (NL) understanding tests, for the set of all answerable queries (Class A+D), the unweighted error rate ranges from 5.9% to 41.7%. For Class A queries, the range is 3.8% to 30.6%, and for Class D, the range is 9.1% to 58.9%.

For Class A+D, the lowest unweighted error rate of 5.9% was obtained by the att1 system. AT&T describes this system as one which "includes a lexical preprocessor which classifies each word of a sentence into lexical/semantic categories, and produces a word lattice. A conceptor processes the word lattice and outputs the most likely segmentation of the original sentence into phrases, for which each phrase corresponds to a single conceptual unit. A template generator then analyzes the phrases obtained by the conceptor and generates a representation of the meaning, and combines it with a similar representation of the context produced by previous sentences".

This year, MADCOW participants agreed to implementation of a McNemar paired-comparison test on both NL and SLS tests. As indicated in Table A15 in the Appendix, the McNemar tests indicate that paired-comparison differences in performance between the AT&T NL system and the CMU and MIT/LCS NL systems were not significant.

Spoken Language System (SLS) Tests
For the Spoken Language System (SLS) tests, two sites (CMU and SRI) submitted SLS results involving the use of rescoring the output of the speech recognition system to select a higher-scoring utterance hypothesis to process to generate the SLS output. Neither of these systems was designated as the "primary" system for scoring and comparative purposes.

For the set of all answerable queries, Class A+D, the unweighted error rate ("UW. Err.") ranges from 8.6% to 55.3%. For Class A queries, the lowest unweighted error rate was 6.5% for the cmu2 system, although that system was not designated as an "official" or comparative system. The next-to-lowest (but "official") Class A error rate (7.0%) was achieved by the att1 system. As in previous years, error rates for context-dependent answerable queries (Class D) were higher, with both the cmu2 and att1 systems achieving error rates of 11.8%, and error rates as high as 71.4% were reported.

For the SLS test results, the McNemar tests indicate that paired-comparison differences in performance between the AT&T system and the two CMU SLS systems are not significant.

4. ACKNOWLEDGEMENTS

The "Hub and Spokes" CSR tests were in large part successfully implemented because of the tireless and effective efforts of

Francis Kubala, as Chair of the ARPA continuous speech recognition Corpus Collection Coordinating Committee (CCCC). The tests relied on speech corpora collected by Denise Danielson and Kate Hunicke-Smith and their colleagues at SRI, and language models developed by Roni Rosenfeld at CMU, using the CMU Statistical Language Modeling Toolkit, working in close collaboration with David Graff and Jack Godfrey at the Linguistic Data Consortium. In the ATIS community, Debbie Dahl served as Chair of the MADCOW group. Kate Hunicke-Smith and others at SRI International were again responsible for annotation of ATIS data and for assisting NIST in the adjudication process following preliminary scoring. It is always a pleasure to acknowledge Kate's thoughtful and cheerful interactions with our group at NIST.

As in previous years, the cooperation of many participants in the ARPA data and test infrastructure -- typically several individuals at each site -- is gratefully acknowledged. Special thanks are due this year to Phil Woodland and Julian Odell at Cambridge University for providing time-marked system output (for the htk1 system) and reference transcription files for our use in evaluating alternative approaches to string alignment and scoring.

Finally, some material in this paper was quoted from System Descriptions provided by participants to NIST and other participants. Similar information will be found in the relevant references. Responsibility for any misquotation or inadvertent misrepresentation is that of the NIST senior author (DSP).

NOTICE

The views expressed in this paper are those of the author(s). The results presented are for local, system-developer-implemented tests. NIST's role in the tests is one of selecting and distributing the test materials, implementing scoring software, and uniformly tabulating the results of the tests. The views of the author(s), and these results, are not to be construed or represented as endorsements of any systems or official findings on the part of NIST, ARPA or the U.S. Government.

REFERENCES

[1] Pallett, D., et al., "1993 Benchmark Tests for the ARPA Spoken Language Program", in Proceedings of the Human Language Technology Workshop, C.J. Weinstein, ed., March 8-11, 1994, Plainsboro, NJ., Morgan Kaufmann Publishers, Inc., ISBN 1-55860-357-3.

[2] Kubala, F., "Design of the 1994 CSR Benchmark Tests", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

[3] Fisher, W. M., et al., "Further Studies in Phonological Scoring", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

[4] Pallett, D., Fiscus, J., and Garofolo, J., "Resource Management Corpus: September 1992 Test Set Benchmark Test Results", in Proceedings of ARPA Microelectronics Technology Office Continuous Speech Recognition Workshop, 21-21 September, 1992, Stanford, CA.

[5] Ebel, W. and Picone, J., "Human Speech Recognition Performance on the 1994 CSR S10 Corpus", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

[6] Dahl, D., (title unknown at publication), presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

[7] Kubala, F., personal communication to D.S. Pallett, January 9, 1995.

[8] Taylor, Barry N. and Kuyatt, Chris E., "Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results", NIST Technical Note 1297, 1994 edition, September 1994.

Note: The following presentations at the ARPA Spoken Language Technology Workshop (and presumably in this Workshop Proceedings) are relevant to the results reported in this paper. They are listed by site, rather than author, to facilitate reference to data tabulated in this paper.

CSR Papers:

AT&T Bell Laboratories
Ljolje, A., et al., "The AT&T 60,000 Word Speech-to-Text System, presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

BBN
Nguyen, L., et al., "The 1994 BBN/BYBLOS Speech Recognition System", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

Zavaliagkos, G., Schwartz, R., and Makhoul, J., "Adaptation Algorithms for BBN's Phonetically Tied Mixtures System", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

Boston University
Ostendorf, M., et al., "The 1994 BU WSJ Benchmark System", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

Cambridge University
Woodland, P., et al., "The Development of the 1994 HTK Large Vocabulary Speech Recognition System", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

Leggetter, C., and Woodland, P., "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

Hochberg, M., et al., "The 1994 ABBOT Hybrid Connectionist-HMM Large-Vocabulary Recognition System", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

Carnegie Mellon University
Chase, L., et al., "Improvements in Language, Lexical, and Phonetic Modeling in Sphinx-II", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

Moreno, P., et al., "Continuous Recognition of Large-Vocabulary Telephone Quality Speech", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

Rudnicky, A., "Language Modeling with Limited Domain Data", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

Rosenfeld, R., "The CMU Statistical Language Modeling Toolkit and its Use in the 1994 ARPA CSR Evaluation", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

Centre de recherche informatique de Montreal (CRIM)
Normandin, Y., et al., "CRIM's November 94 Continuous Speech Recognition System",presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

IBM
Bahl, L., et al., "Performance of the IBM Large Vocabulary Continuous Speech Recognition System on the ARPA NAB News Task", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

Gopinath, R., et al., "Robust Speech Recognition in Noise - Performance of the IBM Continuous speech Recognizer on the ARPA Noise Spoke Task", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

LIMSI
Gauvain, J., Lamel, L., and Adda-Decker, M., "Developments in Continuous Speech Dictation using the ARPA WSJ Task", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

MIT/Lincoln Laboratory
Paul, D., "New Developments in the Lincoln stack-Decoder Based Large-Vocabulary CSR System", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

New York University/BBN
Sekine, S., Sterling, J, and Grishman, R., "NYU/BBN 1994 CSR Evaluation", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

Philips
Dugast. C., et al., "Continuous Speech Recognition Tests and Results for the NAB'94 Corpus", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

SRI
Digilakis, V., et al., "Continuous Speech Dictation on ARPA's North American Business News Domain", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

University of Karlsruhe
Rogina, I., and Waibel, A., "The JANUS Speech Recognizer", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

ATIS Papers:

AT&T
Bocchieri, E., Riccardi, G., and Anantharaman, J., "The 1995 AT&T ATIS Speech Recognizer", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

Levin, E., and Pieraccini, R., "CHRONUS: the New Generation", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

BBN
Miller, S. et al., "Recent Progress in Hidden Understanding Models", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

Stallard, D., "The BBN ATIS4 Dialogue System", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

Carnegie Mellon University
Ward, W., and Issar, S., "The CMU ATIS System", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

MIT/LCS
Zue, V., et al., "The MIT ATIS System: December 1994 Progress Report", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

MITRE
Bayer, S., et al., "Spoken Language Understanding: Report on the MITRE Spoken Language System", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

SRI
Cohen, M., Rivlin, Z., and Bratt, H., "Speech Recognition in the ATIS Domain Using Multiple Knowledge Sources", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

Moore, R., et al., "Combining Linguistic and Statistical Knowledge Sources in Natural-Language Processing for ATIS", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

UNISYS
Dahl, D., et al., "Weakly Supervised Training for Spoken Language Understanding", presented at the ARPA Spoken Language Technology Workshop, 22-25 January, 1995, Austin, TX.

APPENDIX: "BENCHMARK TEST RESULTS"

CSR Test Participants

United States participants in the WSJ-CSR tests included: AT&T Bell Laboratories (ATT), BBN Systems and Technologies (BBN), Boston University (BU), Carnegie Mellon University (CMU), Dragon Systems, IBM T.J. Watson Research Labs (IBM), Massachusetts Institute of Technology's Lincoln Laboratory (MIT/LL), New York University (NYU), and SRI International (SRI).

Foreign participants included two British groups at Cambridge University's Engineering Department, one the developers of the HMM Toolkit (CU-HTK), and another pursuing connectionist approaches (CU-CON), a French group at CNRS-LIMSI (LIMSI), and two German groups, at Karlsruhe University (KU), and at the Philips GmbH Research Laboratories in Aachen (PHIL-TH), in addition to the Canadian Centre de Recherche Informatique de Montreal (CRIM).

This was the first year of participation in the ARPA large vocabulary CSR benchmark tests for Karlsruhe University and CRIM.

The Philips researchers collaborated with others at the Rheinisch Westfaelische Technische Hochschule, also in Aachen.

BU collaborated with BBN, making use of the N-best outputs of a BBN system, using an N-best rescoring formalism, a stochastic segment modelling approach, and the use of several BU and BBN knowledge sources.

NYU also worked jointly with BBN, taking as input the N-best outputs of a BBN system, and making use of both (1) sublanguage, or topic coherence, and (2) syntactic scores to select a new 1-best hypothesis. The goal of the NYU effort is to "determine whether long-range, linguistically based word preferences can be used to enhance speech recognition."

Although not a formal participant in these benchmark tests, Ebel and Picone at Mississippi State University conducted tests involving the use of human transcribers using the S10 data. The objective of these tests was to "benchmark human performance" in recognizing and transcribing noisy speech data.

ATIS Test Participants

Participants in the ATIS tests included: AT&T Bell Laboratories (AT&T), BBN Systems and Technologies (BBN), Carnegie Mellon University (CMU), Massachusetts Institute of Technology's Laboratory for Computer Science (MIT/LCS), the MITRE Corporation, (MITRE), SRI International (SRI), and Unisys (UNISYS).

Unisys collaborated with BBN, using a set of N-best outputs for a BBN ATIS-domain speech recognition system as input for Unisys-developed natural language technology.

MITRE's NL system shared some software originally developed at CMU. A CMU Phoenix parser is used in conjunction with a "simple discourse module, a backend query generation, and a multimodal dialogue system knowledge representation module". This is the first ATIS evaluation in which MITRE has participated.

November 1994 CSR Training and Test Material

Some of the 1994 CSR tests make use of the Linguistic Data Consortium's newly provided language model training material (CSRNAB1) involving North American Business News and a 20K trigram language model developed by researchers at CMU based on this material. The texts and language models are included in a 4-CD-ROM set (LM1) provided by the LDC. Developers also made use of NAB-News development test sets that are identical in design and scope to the 1994 evaluation test sets. The development test sets include material for each hub and spoke test.

November 1994 CSR Test Scoring and Adjudication

During the adjudication period, NIST received 27 email messages containing 460 utterance-specific bug reports (including adjudication requests and comments) pertaining to 218 unique utterances in the H1, H2, and S3 test sets. A few general messages were also received regarding the principles to be followed in adjudicating.

A significant number (151) of the bug reports pertained to requests for the allowance of multiple forms of compound words. These requests were concerned with 49 compound words which occurred in 77 utterances. Most of these compound word problems occurred as a result of inconsistencies in the WSJ and NAB news training data or in improper handling of hyphenation. Of the 49 compound words in question, all but 8 of the requested alternates were allowed, and added to a global map table which was used on all of the reference and hypothesis transcriptions prior to scoring. The remaining 8 compound words were deemed to be homophones and were evaluated on an utterance-by-utterance basis.

The remaining (313) non-compound-word requests pertained to 158 utterances. Of these, 140 requests affecting 84 utterances were changed as requested, 34 requests affecting 25 utterances were partially changed as requested, and 139 requests affecting 99 utterances were denied.

The CSR adjudication process this year was much more intensive than in previous years, and although the requests account for less than 1% of the words in the H1, H2, and S3 test sets, they raised several general issues concerning: (1) compound words, (2) proper name homophones, (3) acoustic ambiguities in connecting words, especially in words ending in "ed" or "s", and (4) acoustic evidence versus syntactic and semantic pressures. These issues merit continuing discussion.

1994 CSR Benchmark Test Results

The text of this appendix is primarily intended to assist the reader in interpreting the data presented in the tabulations of results in this appendix. There is a great deal of detailed data, and many alternative interpretations and contrasts are possible. The tables of this appendix present the results of NIST's "official" scoring and the implementations of paired-comparison statistical significance tests, using formats for the tabulations that have been agreed to by two groups of participants: the Continuous speech recognition Corpus Coordinating Committee (CCCC), and the Multi-site ATIS Data COrpus Working Group (MADCOW).

The reader is referred to other papers in this proceedings, by the developers of the relevant systems, for additional discussion. Note that, in some cases, error rates presented in those papers may differ somewhat from these "official" results, because of local use of unofficial scoring software, pre-adjudicated reference transcriptions, or of systems that differ somewhat from the versions used for the "official" benchmark tests.

Hub 1: Unlimited Vocabulary NAB News Baseline. One goal of this portion of the tests was to document the state-of-the-art in "basic SI (Speaker Independent) performance on clean data".

The "Primary" systems could use "any grammar or acoustic training predating June 16, 1994" and make use of knowledge of session boundaries and utterance order given as side information. In Table A1, these results are shown in the column labelled "P0". Although the H1P0 condition was "the premier test of the entire test suite [2]", note that comparisons involving P0 systems from diffferent sites are complicated by many factors including different acoustic training, language models and procedures for adaptation.

The results for the baseline system, using a static SI test with the 1994 standard 20 K trigram language model, and choice of either of two specified sets of "short-term" or "long term" speakers for system training, are tabulated in the column labelled "Contrast C1" . These restrictions "were designed to permit close comparisons of acoustic recognition capability among all systems [2]."

The column labelled P1 indicates results for the one NYU system (otherwise the same as that site's P0 system) that made use of known article boundaries. The column labelled C2 presents results for systems that were otherwise the same as the site's P0 system, but which incorporated supervised incremental adaptation.

In most cases, data from each site shows on a single line.

The three BU "C1" systems each represent different N-best rescoring formalisms using the BU stochastic segment model recognition system in combination with the BBN Byblos system, using different knowledge sources to re-rank the N-best hypotheses.

The Cambridge University "cu-con1" results were the result of inadvertent operator error, and the "cu-con2" results are what was intended to have been submitted.

There are two sets of results from collaborative research involving NYU and BBN, that involve using the N-best outputs of a BBN system, and making use of sublanguage, or topic coherence scores (nyu1), or syntactic scores (nyu2), to select a new 1-best hypothesis.

The two sets of results from Philips, for H1PO, differ in that the better-performing system ("phil-th2") implements unsupervised speaker adaptation with knowledge of session boundaries. NIST was advised that a "bug" was discovered in the language model for the "phil-th1" P0 system after submission of the results presented in this table.

In this table, and others of this sort in this paper, the results of contrastive comparisons are shown in the boxes labelled "COMPARISONS AND SIGNIFICANCE TESTS". The results of use of the NIST statistical significance tests that have been used in previous tests are also shown.

To illustrate interpretation of some of the tabulated results, note that ATT and BBN achieved reductions in error rate of 23.6% and 14.1%, respectively, for their P0 systems when compared to the C1 baseline systems. In most cases, these reductions were shown to be significant.

Table A2: Table A2 shows a matrix tabulation of the results of cross-site and, in some cases, within-site, paired comparison statistical significance tests for the baseline H1-C1 systems. The number of paired comparisons involved in these significance tests unfortunately necessitates the use of an extremely small font. As in previous years' summary papers, the convention used for these matrix tabulations is to print the name of the system with the lower error rate in the event that the relevant null hypotheses are not shown to be valid, and "same" if the null hypothesis is valid.

Hub 2: The goal of this hub was to "demonstrate SI performance on unlimited vocabulary read speech over long-distance telephone lines". Table A3 documents word error rates achieved by the three sites participating in this second Hub -- CMU, LIMSI, and SRI. No contrastive or significance tests were designated for Hub 2.

Spoke 2: Domain Adaptation As noted elsewhere in this paper, "Spoke 2 involved adaptation to new news topic not found in the training material. CMU was the only site participating in this spoke. The (incomplete) results reported to NIST as of January 17, 1995 do not include the use of topic-specific training, and are thus largely inconclusive." Table A4 indicates results for the P0 and C1 conditions for this spoke. The test material included material on two different news topics -- "China" and [O.J.] "Simpson", and analyses of results on these topics, individually, are shown.

Spoke 3: SI Recognition Outliers (Non-Native Speakers) The stated goal for this spoke was "to evaluate a rapid enrollment speaker adaptation algorithm on difficult speakers (e.g., non-native speakers of American English)". Participants included BBN, CU, and SRI. Test data consisted of read speech from ten speakers, each reading 20 sentence utterances, with the Sennheiser microphone. For each speaker, the 40 "rapid enrollment" utterances were available for use with the "rapid enrollment" speaker adaptation. An additional 160 utterances were available to permit use of a total of three different adaptation sets -- 40 (P0), 100 (C4) and 200 (C5) utterance enrollment sets.

Table A5 presents the results for Spoke 3. The column labelled P0 shows results with 40-utterance rapid enrollment adaptation enabled, error rates of 10.1% to 11.1% were achieved. In contrast, with adaptation disabled, the word error rates range between 20.7% and 26.1%. With use of 200 utterance enrollment, SRI achieved an error rate of 7.8%.

Spoke 4: Incremental Speaker Adaptation. The stated goal for this spoke was "to evaluate an incremental speaker adaptation algorithm". Three sites participated: BBN, CU and Dragon. In this spoke, there were only four test speakers, with 100 sentence utterances for each. NIST's scoring was done on four successive 25-sentence utterance blocks (i.e., utterances 1-25, 26-50, 51-75, and 76+).

Table A6 presents the results for Spoke 4. The lowest error rates for this spoke were obtained by the Cambridge University's HTK system, with word error rates for the P0 condition (with incremental unsupervised adaptation enabled) ranging from 5.0% to 7.8% for the four 25-utterance blocks. For the BBN system, corresponding word error rates range from 9.0% to 11.8%. For the Dragon results, the range in word error rates for the P0 condition is from 9.0% to 11.1%.

Spoke 5: "Microphone Independence". The stated goal of this spoke was to "evaluate an unsupervised channel compensation algorithm". The different "channels" in this case were different microphones -- the twenty speakers in this test set used ten different (unknown) microphones. Similar, but not identical, microphones had been incorporated in training and development test material.

CMU was the sole participant in this spoke.

The "cmu2" system used for this spoke incorporated a new algorithm called "RATZ (snR-dependent gAussian-based cepstTral normaliZation)", used to compensate for the effects of recording speech through an unknown microphone. The "cmu4" system used "a new version of the CDCN (Codebook-Dependent Cepstral Normalization) used to compensate for the combined effects of additive noise and unknown channel distortion.

Table A7 presents the results for Spoke 5. With unsupervised channel compensation enabled, the CMU systems achieved error rates of 9.9% and 9.7%, in contrast to 12.4% with compensation disabled -- approximately a 20% reduction in word error rate.

Spoke 9: Spontaneous WSJ Dictation. The stated goal of this spoke was to "improve basic performance on spontaneous dictation-style speech". There were 10 speakers, each dictating 20 spontaneous Wall Street Journal-like sentence utterances, and using the Sennheiser microphone.

BBN [13] was the sole participant in this spoke.

Table A8 presents the results for Spoke 9. Using the same system as used for the P0 condition in Hub 1 (which achieved a word error rate of 10.2% on the Hub 1 test data), a word error rate of 14.2% was achieved on the S9 data.

Spoke 10: Noisy Channel. The stated goal of this spoke was to evaluate compensation on data corrupted with additive noise. There were 10 speakers, each speaking 10 utterances, with the 100 utterance test set presented when additively combined with recorded automobile interior noise at three unknown (a priori) A-weighted SNRs: (22 dB, 16 dB, and 10 dB). Table A9 presents the results for Spoke 10. Results are presented for each SNR condition with compensation enabled (P0), and disabled (C1), and for the case of compensation disabled for the "clean" data without any additive noise. Participants included the developers of the Cambridge University HMM Toolkit, IBM, and SRI.

ATIS November 1994 Test Material

The final, adjudicated set of test material consisted of 981 test utterances and was collected at 5 sites -- BBN, CMU, MIT, NIST and SRI. As in previous years, it was selected by NIST staff from set-aside material previously collected within the MADCOW community. The test set was selected so as to balance the number of utterances per data collection site (~200 utterances per site.) Data collected at NIST made use of BBN- and SRI-developed ATIS systems. Because of differences in the scenarios and data collection systems used at the different collection sites, it was not possible to balance the test set for number of subjects or the difficulty of scenarios per collection site. No "pre-filtering" of the test data was performed except to attempt to exclude subject-scenarios with mostly repetitive queries. The ATIS test material was released in November, 1994.

1994 ATIS Scoring and Adjudication

During the adjudication period for the December 1994 ATIS evaluation, a total of 125 requests for adjudication (bug reports) were filed with NIST. Of these, 19 reported on problems that were already reported on by others, leaving 106 net problems.

A new procedure for cooperative adjudication work was used this year, with good results: SRI remotely logged on to a NIST computer so that both NIST and SRI could add comments to one copy of the bug report files. Manual semaphoring kept the two sites from trying to edit the same file simultaneously. Productivity was increased and the likelihood of manual errors decreased by the elimination of copying and duplication.

These bug reports were divided into problems with transcription and problems with interpretation; NIST initially tackled the transcription problems while SRI took the interpretation ones. About half fell into each category. After preliminary decisions had been made on them all, each of the two adjudicating sites reviewed the other's analyses and decisions; then disagreements were discussed and final decisions made, with NIST having the final say. Ultimately there was only one bug report on which the judgement of NIST and SRI differed.

The types of problems reported this year have all been seen before. The most interesting of these, raising questions of just what aspects of prior context should be carried forward into an interpretation, seems to us rather murky and deserving of more empirical research.

1994 ATIS Benchmark Test Results

SPontaneous speech RECognition (SPREC) Tests.
Table A10 presents the results for the SPREC tests for all systems and subsets of the ATIS test data, using the Sennheiser close-talking microphone. For the case of the subset of all answerable queries, Class A+D, the word error rates ranged from 1.9% to 14.1%. Two sets of results were submitted by SRI, one designated as a "primary" system (sri3), and one additional set for an N-best system (sri4). Both of these systems performed well.

Table A11 presents a matrix tabulation of the ATIS SPREC results for the Class A+D subset. The overall word error rate across all tested systems for the data from the several collecting sites ("Overall Totals" row along the bottom of the Table) ranges from 2.0% for the NIST-BBN-system collected data to 9.1% for the SRI-collected data, reflecting differences in subject populations and other factors.

Table A12 presents the results, in matrix form, of the application of 4 paired-comparison significance tests for the SPREC systems for the Class A+D subset. For the SRI N-best system (sri4), for these paired comparison significance tests, there are a total of 36 paired-comparison significance tests (9 systems X 4 significance tests), and all but 2 of these indicated significant SPREC test differences with lower error rates for the sri4 system.

Natural Language (NL) Understanding Tests.
Table A13 presents a tabulation of the results for the NL tests for all systems and all sets of "answerable" ATIS queries, Class A+D, Class A and Class D. For the set of all answerable queries, Class A+D, the unweighted error rate ("UW. Err.") ranges from 5.9% to 41.7%. For Class A queries, the range is 3.8% to 30.6%, and for Class D, the range is 9.1% to 58.9%.

For Class A+D, the lowest unweighted error rate of 5.9% was obtained by the att1 system.

Table A14 presents a matrix tabulation of the official NL results for the several subsets of test material, for Class A+D. There is some indication of varying degrees of difficulty presented by the different subsets of data from the different sites, subject-scenarios, and subject populations: note that the unweighted error rates reported in the "Overall Totals" row ranges from 9.1% to 19.6%.

This year, MADCOW participants agreed to implementation of a McNemar paired-comparison test on both NL and SLS tests (on whether or not queries in Class A+D were correctly answered). Table A15 shows the results of those tests, for NL results for the Class A+D queries. Comparisons involving the att1 system indicate that differences in the percent correctly answered for both the cmu1 and the mit_lcs1 systems were not statistically significant, although comparisons of the performance of the att1 involving other systems were significant.

Spoken Language System (SLS) Understanding Tests.
Table A16 presents a tabulation of the results for the SLS tests for all systems and all sets of "answerable" ATIS queries, Class A+D, Class A and Class D. Note that two sites continued to use the "No_Answer" option, although other sites abandoned use of this option since the use of unweighted error measures offers no strategic benefit for use.

CMU and SRI each submitted two sets of results, and in each case, one of the two systems made use of N-best utterance hypothesis lists provided by the speech recognition module, and implemented forms of rescoring before selecting the 1-best output for processing by the NL module. The systems using rescoring are the cmu2 and sri2 systems.

For the set of all answerable queries, Class A+D, the unweighted error rate ("UW. Err.") ranges from 8.6% to 55.3%. For Class A queries, the lowest unweighted error rate was 6.5% for the "unofficial" cmu2 system using rescoring, and the next-to-lowest error rate was 7.0% for the att1 system. For Class D, the range is 11.8% to 71.4%.

Table A17 presents a matrix tabulation of the SLS results for the several subsets of Class (A+D) test material from different sites. This year, there is little evidence of "local adaptation" to locally collected data.

Table A18 shows the results of the McNemar test applied to the SLS results. Comparisons involving the att1 system indicate that differences in the percent correctly answered for the paired-comparisons involving the two CMU systems were not statistically significant, although comparisons involving other systems were significant.

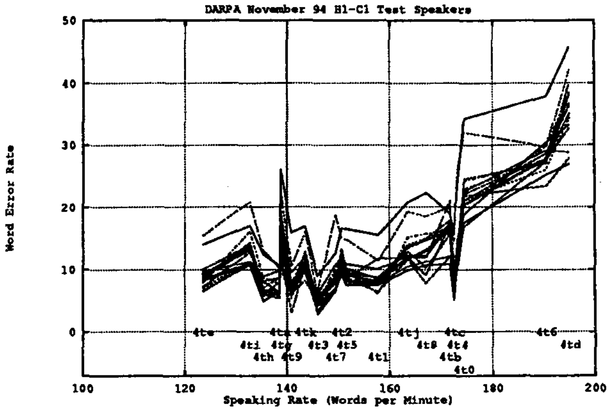

Figure 1 Word error rates for Hub1 C1 speakers vs. speaking rate

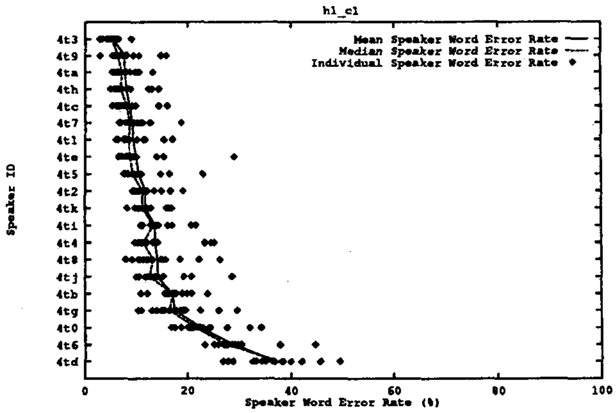

Figure 2 Variablity in error rate among the twenty Hub1 C1 systems for the twenty speakers in the Hub 1 test set

Nov 94 Hub and Spoke CSR Evaluation
Hub 1: Unlimited Vocab. NAB News Baseline

GOAL:	Improve basic SI performance on clean data
DATA:	20 A speakers * 15 utts = 300 utts from several sources of North American Business (NAB) News, collected from June 16, 1994 to July 15, 1994, Sennheiser mic.

Primary and Contrast Conditions

P0-IV	Same as H1-P0, In vocabulary subset
P0-OOV	Same as H1-P0, Out of vocabulary subset
C1-IV	Same as H1-C1, In vocabulary subset
C1-OOV	Same as H1-C1, Out of vocabulary subset
SIDE INFO:	Session boundaries and utterance order are known for H1-P0 only.

	Hub 1 P0 In-Vocab.		Hub 1 P0 Out-of-Vocab.		Hub 1 C1 In-Vocab.		Hub 1 C1 Out-of-Vocab.	
System	W.E. (%)	U.E.	W.E. (%)	U.E.	W.E. (%)	U.E.	W.E. (%)	U.E.
att1	7.5	57.9	13.0	84.1	8.5	61.6	18.7	95.2
bbn1	8.2	61.1	12.7	83.3	8.0	58.4	16.8	96.8
bu1	9.0	62.1	13.7	88.1	9.1	62.1	17.5	97.6
bu2	8.8	65.8	13.6	90.5	9.1	66.3	17.8	96.8
bu3	9.2	70.5	13.8	92.9	9.7	73.7	18.5	96.8
cmu1	8.6	65.8	13.6	84.9	9.0	64.7	19.4	97.6
crim1	16.0	83.2	25.4	99.2	16.0	83.2	25.4	99.2
cu-con1	12.0	71.6	17.8	89.7	9.6	66.8	19.7	96.8
cu-con2	10.0	68.4	15.3	88.1	9.6	66.8	19.7	96.8
cu-htk1	5.3	50.5	9.5	76.2	6.4	58.9	15.6	96.0
dragon1	8.3	65.8	12.8	87.3	8.8	65.3	18.8	96.0
ibm1	6.2	52.6	11.5	85.7	6.5	56.3	16.9	96.8
ku1	18.2	81.1	28.5	99.2	18.2	81.1	28.5	99.2
limsi1	7.0	58.9	11.9	79.4	7.4	58.4	17.8	97.6
mit-ll1	15.0	79.5	20.3	93.7	14.8	78.4	24.1	97.6
nyu1	9.1	64.2	13.3	88.1	9.1	64.2	17.8	96.8
nyu2	9.1	63.7	13.4	88.1	9.1	64.2	17.8	96.8
phil-th1	9.1	69.5	14.9	93.7	9.3	70.0	18.5	97.6
phil-th2	8.1	65.3	13.6	89.7	9.3	70.0	18.5	97.6
sri1	8.2	62.6	12.9	84.9	8.1	61.6	17.2	97.6

COMPARISONS AND SIGNIFICANCE TESTS

	Test Comp.	% Reduct. W.E.	Significance Tests: Sign	Wilcoxon
att1	P0-IV:C1-IV	11.8%	P0-IV	P0-IV
bbn1	P0-IV:C1-IV	-2.7%	same	same
bu1	P0-IV:C1-IV	1.5%	same	same
bu2	P0-IV:C1-IV	3.4%	same	same
bu3	P0-IV:C1-IV	4.6%	same	same
cmu1	P0-IV:C1-IV	4.6%	same	same
crim1	P0-IV:C1-IV	0.0%	same	same
cu-con1	P0-IV:C1-IV	-25.3%	C1-IV	C1-IV
cu-con2	P0-IV:C1-IV	-3.9%	same	same
cu-htk1	P0-IV:C1-IV	16.0%	same	same
dragon1	P0-IV:C1-IV	5.3%	same	same
ibm1	P0-IV:C1-IV	4.4%	same	same
ku1	P0-IV:C1-IV	0.0%	same	same
limsi1	P0-IV:C1-IV	5.1%	same	same
mit-ll1	P0-IV:C1-IV	-1.8%	same	same
nyu1	P0-IV:C1-IV	0.3%	same	same
nyu2	P0-IV:C1-IV	0.7%	same	same
phil-th1	P0-IV:C1-IV	1.4%	same	same
phil-th2	P0-IV:C1-IV	12.2%	P0-IV	P0-IV
sri1	P0-IV:C1-IV	-0.3%	same	same

	Test Comp.	% Reduct. W.E.	Significance Tests: Sign	Wilcoxon
att1	P0-OOV:C1-OOV	30.3%	P0-OOV	P0-OOV
bbn1	P0-OOV:C1-OOV	24.0%	P0-OOV	P0-OOV
bu1	P0-OOV:C1-OOV	21.8%	P0-OOV	P0-OOV
bu2	P0-OOV:C1-OOV	23.9%	P0-OOV	P0-OOV
bu3	P0-OOV:C1-OOV	25.3%	P0-OOV	P0-OOV
cmu1	P0-OOV:C1-OOV	29.7%	P0-OOV	P0-OOV
crim1	P0-OOV:C1-OOV	0.0%	same	same
cu-con1	P0-OOV:C1-OOV	9.3%	same	P0-OOV
cu-con2	P0-OOV:C1-OOV	22.3%	P0-OOV	P0-OOV
cu-htk1	P0-OOV:C1-OOV	39.4%	P0-OOV	P0-OOV
dragon1	P0-OOV:C1-OOV	31.9%	P0-OOV	P0-OOV
ibm1	P0-OOV:C1-OOV	31.9%	P0-OOV	P0-OOV
ku1	P0-OOV:C1-OOV	0.0%	same	same
limsi1	P0-OOV:C1-OOV	33.1%	P0-OOV	P0-OOV
mit-ll1	P0-OOV:C1-OOV	15.9%	P0-OOV	P0-OOV
nyu1	P0-OOV:C1-OOV	25.5%	P0-OOV	P0-OOV
nyu2	P0-OOV:C1-OOV	24.7%	P0-OOV	P0-OOV
phil-th1	P0-OOV:C1-OOV	19.8%	P0-OOV	P0-OOV
phil-th2	P0-OOV:C1-OOV	26.5%	P0-OOV	P0-OOV
sri1	P0-OOV:C1-OOV	24.9%	P0-OOV	P0-OOV

	Test Comp.	% Reduct. W.E.	Significance Tests: Sign	Wilcoxon
att1	C1-IV:C1-OOV	54.6%	C1-IV	C1-IV
bbn1	C1-IV:C1-OOV	52.4%	C1-IV	C1-IV
bu1	C1-IV:C1-OOV	48.1%	C1-IV	C1-IV
bu2	C1-IV:C1-OOV	48.9%	C1-IV	C1-IV
bu3	C1-IV:C1-OOV	47.6%	C1-IV	C1-IV
cmu1	C1-IV:C1-OOV	53.5%	C1-IV	C1-IV
crim1	C1-IV:C1-OOV	37.1%	C1-IV	C1-IV
cu-con1	C1-IV:C1-OOV	51.1%	C1-IV	C1-IV
cu-con2	C1-IV:C1-OOV	51.1%	C1-IV	C1-IV
cu-htk1	C1-IV:C1-OOV	59.2%	C1-IV	C1-IV
dragon1	C1-IV:C1-OOV	53.2%	C1-IV	C1-IV
ibm1	C1-IV:C1-OOV	61.4%	C1-IV	C1-IV
ku1	C1-IV:C1-OOV	36.3%	C1-IV	C1-IV
limsi1	C1-IV:C1-OOV	58.3%	C1-IV	C1-IV
mit-ll1	C1-IV:C1-OOV	38.9%	C1-IV	C1-IV
nyu1	C1-IV:C1-OOV	48.8%	C1-IV	C1-IV
nyu2	C1-IV:C1-OOV	48.8%	C1-IV	C1-IV
phil-th1	C1-IV:C1-OOV	50.1%	C1-IV	C1-IV
phil-th2	C1-IV:C1-OOV	50.1%	C1-IV	C1-IV
sri1	C1-IV:C1-OOV	52.8%	C1-IV	C1-IV

Table 1 Error rates for IV-containing and OOV-Containing subsets of Hub 1

System	IV Subset Err Rate(%)	IV Induced Errors in OOV Subset	IV-Induced +OOV Errors	Total OOV Errors	Additional Errors	Additional Error Ratio
att1	8.5	294	488	684	196	2.0
bbn1	8.0	277	471	613	142	1.7
bu1	9.1	315	509	641	132	1.7
bu2	9.1	315	509	652	143	1.7
bu3	9.7	336	530	675	145	1.7
cmu1	9.0	312	506	710	204	2.1
crim1	16.0	554	748	930	182	1.9
cu-con1	9.6	333	527	719	192	2.0
cu-con2	9.6	333	527	719	192	2.0
cu-htk1	6.4	222	416	571	155	1.8
dragon1	8.8	305	499	686	187	2.0
ibm1	6.5	225	419	617	198	2.0
ku1	18.2	631	825	1043	218	2.1
limsi1	7.4	256	450	652	202	2.0
mit-ll1	14.8	513	707	883	176	1.9
nyu1	9.1	315	509	652	143	1.7
nyu2	9.1	315	509	652	143	1.7
phil-th1	9.3	322	516	678	162	1.8
phil-th2	9.3	322	516	678	162	1.8
sri1	8.1	281	475	630	155	1.8

Table 2 Estimates of the number of word errors per OOV-word occurrence

```
                                 NON-OOV WORDS
          Word must contain 2 non-oov words on both sides to be considered
                  #     WRD(-2)     #    WRD(-1)    #  WRD    #    WRD(+1)    #     WRD(+2)     #
#################################################################################################
                  # C   S   D    I  # C   S   D     #  ADJ    # C   S   D     # C   S   D    I  #
             Num  # O   U   E    N  # O   U   E     #  WRD    # O   U   E     # O   U   E    N  #
     Word    Occ  # R   B   L    S  # R   B   L     #  ERR    # R   B   L     # R   B   L    S  #
#################################################################################################
            6230  #5788 344  98   37#5797 336  97   #   31    #5813 321  96   #5801 329 100  39 #
------------------#-----------------#---------------#---------#---------------#-----------------#
          Correct # 92.9%           # 93.0%         # 93.3%   # 93.3%         # 93.1%           #
     Substitution #    5.5%         #    5.4%       #  5.3%   #      5.2%     #     5.3%        #
         Deletion #         1.6%    #           1.6%#  1.4%   #           1.5%#          1.6%   #
        Insertion #             0.6%#               #  0.5%   #               #             0.6%#
#################################################################################################
                  #     WRD(-2)     #    WRD(-1)    #  WRD    #    WRD(+1)    #     WRD(+2)     #

                                    OOV WORDS
            OOV split must contain 2 words on both sides to be considered
                  #     OOV(-2)     #    OOV(-1)    #  OOV    #    OOV(+1)    #     OOV(+2)     #
#################################################################################################
                  # C   S   D    I  # C   S   D     #  ADJ    # C   S   D     # C   S   D    I  #
             Num  # O   U   E    N  # O   U   E     #  OOV    # O   U   E     # O   U   E    N  #
     Word    Occ  # R   B   L    S  # R   B   L     #  ERR    # R   B   L     # R   B   L    S  #
#################################################################################################
             151  # 136  8   6    9 # 110  35   6   #   252   # 112  33   6   # 129  18   4  10 #
------------------#-----------------#---------------#---------#---------------#-----------------#
          Correct # 90.1%           # 72.8%         #   0.0%  # 74.2%         # 85.4%           #
     Substitution #    5.3%         #    23.2%      # 100.0%  #     21.9%     #     11.9%       #
         Deletion #         4.0%    #           4.0%#   0.0%  #           4.0%#          2.6%   #
        Insertion #             6.0%#               #  66.9%  #               #             6.6%#
#################################################################################################
                  #     OOV(-2)     #    OOV(-1)    #  OOV    #    OOV(+1)    #     OOV(+2)     #

                                           DIFFERENCES
                  ###############################################################################
          Correct #     -2.8%       #    -20.2%     # -93.3%  #    -19.1%     #      -7.7%      #  -143.1%
     Substitution #      0.2%       #     17.8%     #  94.7%  #     16.7%     #       6.6%      #   136.0%
         Deletion #      2.4%       #      2.4%     #  -1.4%  #      2.5%     #       1.0%      #     6.9%
                  #                 #               #         #               #                 #
        Insertion #      5.4%       #               #  66.4%  #               #       6.0%      #    78.3%
                  ###############################################################################
```

Table 3 Error rates in the regions surrounding IV and OOV word Occurrences

Nov 94 Hub and Spoke CSR Evaluation
Hub 1: Unlimited Vocab. NAB News Baseline

GOAL: Improve basic SI performance on clean data

DATA: 20 A speakers * 15 utts = 300 utts from several sources of North American Business (NAB) News, collected from June 16, 1994 to July 15, 1994, Sennheiser mic.

Primary and Contrast Conditions

P0 (req) any grammar or acoustic training, predating June 16, 1994, session boundaries and utterance order given as side information.

C1 (req) Static SI test with the 1994 standard 20K trigram LM and choice of either short-term or long-term speakers

P1 (opt) Same as H1 P0 system 'augmented with knowledge of article boundaries'

C2 (opt) Same as P0 with supervised incremental adaptation

SIDE INFO: Session boundaries and utterance order are known for H1-P0 only.

	Primary P0	Contrast C1	Primary P1	Contrast C2
System	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)
att1	10.0	13.0		
bbn1	10.2	11.9		
bu1	11.1	12.9 #		
bu2	10.9	13.0		10.9
bu3	11.3	13.6		11.2
cmu1	10.9	13.7		
crim1	20.2 *	20.2		
cu-con1	14.6 +	14.1		
cu-con2	12.4	14.1		
cu-htk1	7.2	10.5		
dragon1	10.3	13.2		
ibm1	8.6	11.1		
ku1	22.8 *	22.8		
limsi1	9.2	12.1		
mit-ll1	17.4	19.0		
nyu1	11.0	13.0	10.6	10.7
nyu2	11.0	13.0		
phil-th1	11.7	13.4		
phil-th2	10.6	13.4 @		
sri1	10.3	12.2		

* (C1 Results) + (Operator Error) @ (phil-th1 C1 Results)
(Late: Revised Submission to Correct Reporting Error)

COMPARISONS AND SIGNIFICANCE TESTS

	Test	% Reduct.	Significance Tests:			
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
att1	P0:C1	23.6%	P0	P0	P0	P0
bbn1	P0:C1	14.1%	same	P0	P0	P0
bu1	P0:C1	13.9% #	P0	P0	P0	P0
bu2	P0:C1	16.0%	same	P0	P0	P0
bu3	P0:C1	17.2%	same	P0	P0	P0
cmu1	P0:C1	20.6%	same	P0	P0	P0
crim1	P0:C1	0.0% *	same	same	same	same
cu-con1	P0:C1	-3.7% +	same	same	same	same
cu-con2	P0:C1	12.4%	same	P0	same	P0
cu-htk1	P0:C1	31.6%	P0	P0	P0	P0
dragon1	P0:C1	22.2%	same	P0	P0	P0
ibm1	P0:C1	23.0%	P0	P0	P0	P0
ku1	P0:C1	0.0% *	same	same	same	same
limsi1	P0:C1	23.6%	P0	P0	P0	P0
mit-ll1	P0:C1	8.3%	same	P0	P0	P0
nyu1	P0:C1	15.7%	same	P0	P0	P0
nyu2	P0:C1	15.4%	same	P0	P0	P0
phil-th1	P0:C1	12.8%	same	P0	P0	P0
phil-th2	P0:C1	21.1% @	P0	P0	P0	P0
sri1	P0:C1	15.6%	P0	P0	P0	P0

	Test	% Reduct.	Significance Tests:			
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
nyu1	P1:P0	3.0%	same	same	same	same

	Test	% Reduct.	Significance Tests:			
	Comp.	W.E.	McN	MAPSSWE	Sign	Wilcoxon
bu2	C2:P0	0.1%	same	same	same	same
bu3	C2:P0	0.9%	same	same	same	same
nyu1	C2:P0	2.2%	same	same	same	same

Table A1 Hub 1 Results

Composite Report of All Significance Tests
For the h1_c1 Test

Test Name	Abbrev.
Matched Pair Sentence Segment (Word Error) Test	MP
Signed Paired Comparison (Speaker Word Accuracy Rate (%)) Test	SI
Wilcoxon Signed Rank (Speaker Word Accuracy Rate (%)) Test	WI
McNemar (Sentence Error) Test	MN

	att1-h1_c1	bbn1-h1_c1	bu1-h1_c1	bu2-h1_c1	bu3-h1_c1	cmu1-h1_c1	esm1-h1_c1	cu-con1-h1_c1	cu-con2-h1_c1	cu-htk1-h1_c1
att1-h1_c1		MP bbn1-h1_c1 SI same WI bbn1-h1_c1 MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN att1-h1_c1	MP same SI same WI same MN same	MP att1-h1_c1 SI att1-h1_c1 WI att1-h1_c1 MN att1-h1_c1	MP att1-h1_c1 SI same WI same MN same	MP att1-h1_c1 SI same WI same MN same	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN same
bbn1-h1_c1			MP bbn1-h1_c1 SI same WI bbn1-h1_c1 MN same	MP bbn1-h1_c1 SI bbn1-h1_c1 WI bbn1-h1_c1 MN bbn1-h1_c1	MP bbn1-h1_c1 SI bbn1-h1_c1 WI bbn1-h1_c1 MN bbn1-h1_c1	MP bbn1-h1_c1 SI bbn1-h1_c1 WI bbn1-h1_c1 MN same	MP bbn1-h1_c1 SI bbn1-h1_c1 WI bbn1-h1_c1 MN bbn1-h1_c1	MP bbn1-h1_c1 SI bbn1-h1_c1 WI bbn1-h1_c1 MN bbn1-h1_c1	MP bbn1-h1_c1 SI bbn1-h1_c1 WI bbn1-h1_c1 MN bbn1-h1_c1	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN same
bu1-h1_c1				MP same SI same WI same MN same	MP bu1-h1_c1 SI bu1-h1_c1 WI bu1-h1_c1 MN bu1-h1_c1	MP same SI same WI bu1-h1_c1 MN same	MP bu1-h1_c1 SI bu1-h1_c1 WI bu1-h1_c1 MN bu1-h1_c1	MP bu1-h1_c1 SI bu1-h1_c1 WI bu1-h1_c1 MN same	MP bu1-h1_c1 SI bu1-h1_c1 WI bu1-h1_c1 MN same	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN same
bu2-h1_c1					MP bu2-h1_c1 SI bu2-h1_c1 WI bu2-h1_c1 MN bu2-h1_c1	MP same SI same WI same MN same	MP bu2-h1_c1 SI bu2-h1_c1 WI bu2-h1_c1 MN bu2-h1_c1	MP bu2-h1_c1 SI same WI bu2-h1_c1 MN same	MP bu2-h1_c1 SI same WI bu2-h1_c1 MN same	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN cu-htk1-h1_c1
bu3-h1_c1						MP same SI same WI same MN cmu1-h1_c1	MP bu3-h1_c1 SI bu3-h1_c1 WI bu3-h1_c1 MN bu3-h1_c1	MP same SI same WI same MN same	MP same SI same WI same MN same	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN cu-htk1-h1_c1
cmu1-h1_c1							MP cmu1-h1_c1 SI cmu1-h1_c1 WI cmu1-h1_c1 MN cmu1-h1_c1	MP same SI same WI same MN same	MP same SI same WI same MN same	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN same
esm1-h1_c1								MP cu-con1-h1_c1 SI cu-con1-h1_c1 WI cu-con1-h1_c1 MN cu-con1-h1_c1	MP cu-con2-h1_c1 SI cu-con2-h1_c1 WI cu-con2-h1_c1 MN cu-con2-h1_c1	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN cu-htk1-h1_c1
cu-con1-h1_c1									MP same SI same WI same MN same	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN cu-htk1-h1_c1
cu-con2-h1_c1										MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN cu-htk1-h1_c1
cu-htk1-h1_c1										
dragon1-h1_c1										
ibm1-h1_c1										
ku1-h1_c1										
limsi1-h1_c1										
mit-ll1-h1_c1										
nyu1-h1_c1										
nyu2-h1_c1										
phil-th1-h1_c1										
phil-th2-h1_c1										
sri1-h1_c1										

	dragon1-h1_c1	ibm1-h1_c1	ku1-h1_c1	limsi1-h1_c1	mit-ll1-h1_c1	nyu1-h1_c1	nyu2-h1_c1	phil-th1-h1_c1	phil-th2-h1_c1	sri1-h1_c1
att1-h1_c1	MP same SI same WI same MN same	MP ibm1-h1_c1 SI same WI ibm1-h1_c1 MN same	MP att1-h1_c1 SI att1-h1_c1 WI att1-h1_c1 MN att1-h1_c1	MP limsi1-h1_c1 SI same WI limsi1-h1_c1 MN same	MP att1-h1_c1 SI att1-h1_c1 WI att1-h1_c1 MN att1-h1_c1	MP same SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN att1-h1_c1	MP same SI same WI same MN att1-h1_c1	MP sri1-h1_c1 SI sri1-h1_c1 WI sri1-h1_c1 MN same
bbn1-h1_c1	MP bbn1-h1_c1 SI bbn1-h1_c1 WI bbn1-h1_c1 MN same	MP ibm1-h1_c1 SI same WI same MN same	MP bbn1-h1_c1 SI bbn1-h1_c1 WI bbn1-h1_c1 MN bbn1-h1_c1	MP same SI same WI same MN same	MP bbn1-h1_c1 SI bbn1-h1_c1 WI bbn1-h1_c1 MN bbn1-h1_c1	MP bbn1-h1_c1 SI bbn1-h1_c1 WI bbn1-h1_c1 MN bbn1-h1_c1	MP bbn1-h1_c1 SI bbn1-h1_c1 WI bbn1-h1_c1 MN bbn1-h1_c1	MP bbn1-h1_c1 SI bbn1-h1_c1 WI bbn1-h1_c1 MN bbn1-h1_c1	MP bbn1-h1_c1 SI bbn1-h1_c1 WI bbn1-h1_c1 MN bbn1-h1_c1	MP same SI same WI same MN same
bu1-h1_c1	MP same SI same WI same MN same	MP ibm1-h1_c1 SI ibm1-h1_c1 WI ibm1-h1_c1 MN same	MP bu1-h1_c1 SI bu1-h1_c1 WI bu1-h1_c1 MN bu1-h1_c1	MP same SI same WI same MN same	MP bu1-h1_c1 SI bu1-h1_c1 WI bu1-h1_c1 MN bu1-h1_c1	MP same SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN bu1-h1_c1	MP same SI same WI same MN bu1-h1_c1	MP same SI same WI same MN same
bu2-h1_c1	MP same SI same WI same MN same	MP ibm1-h1_c1 SI same WI ibm1-h1_c1 MN ibm1-h1_c1	MP bu2-h1_c1 SI bu2-h1_c1 WI bu2-h1_c1 MN bu2-h1_c1	MP limsi1-h1_c1 SI same WI same MN limsi1-h1_c1	MP bu2-h1_c1 SI bu2-h1_c1 WI bu2-h1_c1 MN bu2-h1_c1	MP same SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP sri1-h1_c1 SI same WI same MN same
bu3-h1_c1	MP same SI same WI same MN dragon1-h1_c1	MP ibm1-h1_c1 SI ibm1-h1_c1 WI ibm1-h1_c1 MN ibm1-h1_c1	MP bu3-h1_c1 SI bu3-h1_c1 WI bu3-h1_c1 MN bu3-h1_c1	MP limsi1-h1_c1 SI limsi1-h1_c1 WI limsi1-h1_c1 MN limsi1-h1_c1	MP bu3-h1_c1 SI bu3-h1_c1 WI bu3-h1_c1 MN same	MP nyu1-h1_c1 SI same WI same MN nyu1-h1_c1	MP nyu2-h1_c1 SI same WI same MN nyu2-h1_c1	MP same SI same WI same MN same	MP same SI same WI same MN same	MP sri1-h1_c1 SI sri1-h1_c1 WI sri1-h1_c1 MN sri1-h1_c1
cmu1-h1_c1	MP same SI same WI same MN same	MP ibm1-h1_c1 SI ibm1-h1_c1 WI ibm1-h1_c1 MN ibm1-h1_c1	MP cmu1-h1_c1 SI cmu1-h1_c1 WI cmu1-h1_c1 MN cmu1-h1_c1	MP limsi1-h1_c1 SI same WI limsi1-h1_c1 MN same	MP cmu1-h1_c1 SI cmu1-h1_c1 WI cmu1-h1_c1 MN cmu1-h1_c1	MP same SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP sri1-h1_c1 SI same WI sri1-h1_c1 MN same
esm1-h1_c1	MP dragon1-h1_c1 SI dragon1-h1_c1 WI dragon1-h1_c1 MN dragon1-h1_c1	MP ibm1-h1_c1 SI ibm1-h1_c1 WI ibm1-h1_c1 MN ibm1-h1_c1	MP esm1-h1_c1 SI same WI esm1-h1_c1 MN same	MP limsi1-h1_c1 SI limsi1-h1_c1 WI limsi1-h1_c1 MN limsi1-h1_c1	MP mit-ll1-h1_c1 SI same WI same MN same	MP nyu1-h1_c1 SI nyu1-h1_c1 WI nyu1-h1_c1 MN nyu1-h1_c1	MP nyu2-h1_c1 SI nyu2-h1_c1 WI nyu2-h1_c1 MN nyu2-h1_c1	MP phil-th1-h1_c1 SI phil-th1-h1_c1 WI phil-th1-h1_c1 MN phil-th1-h1_c1	MP phil-th2-h1_c1 SI phil-th2-h1_c1 WI phil-th2-h1_c1 MN phil-th2-h1_c1	MP sri1-h1_c1 SI sri1-h1_c1 WI sri1-h1_c1 MN sri1-h1_c1
cu-con1-h1_c1	MP dragon1-h1_c1 SI same WI same MN same	MP ibm1-h1_c1 SI ibm1-h1_c1 WI ibm1-h1_c1 MN ibm1-h1_c1	MP cu-con1-h1_c1 SI cu-con1-h1_c1 WI cu-con1-h1_c1 MN cu-con1-h1_c1	MP limsi1-h1_c1 SI limsi1-h1_c1 WI limsi1-h1_c1 MN limsi1-h1_c1	MP cu-con1-h1_c1 SI cu-con1-h1_c1 WI cu-con1-h1_c1 MN cu-con1-h1_c1	MP nyu1-h1_c1 SI nyu1-h1_c1 WI nyu1-h1_c1 MN same	MP nyu2-h1_c1 SI nyu2-h1_c1 WI nyu2-h1_c1 MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP sri1-h1_c1 SI sri1-h1_c1 WI sri1-h1_c1 MN same
cu-con2-h1_c1	MP dragon1-h1_c1 SI same WI same MN same	MP ibm1-h1_c1 SI ibm1-h1_c1 WI ibm1-h1_c1 MN ibm1-h1_c1	MP cu-con2-h1_c1 SI cu-con2-h1_c1 WI cu-con2-h1_c1 MN cu-con2-h1_c1	MP limsi1-h1_c1 SI limsi1-h1_c1 WI limsi1-h1_c1 MN limsi1-h1_c1	MP cu-con2-h1_c1 SI cu-con2-h1_c1 WI cu-con2-h1_c1 MN cu-con2-h1_c1	MP nyu1-h1_c1 SI nyu1-h1_c1 WI nyu1-h1_c1 MN same	MP nyu2-h1_c1 SI nyu2-h1_c1 WI nyu2-h1_c1 MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP sri1-h1_c1 SI sri1-h1_c1 WI sri1-h1_c1 MN same
cu-htk1-h1_c1	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN same	MP same SI cu-htk1-h1_c1 WI same MN same	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN cu-htk1-h1_c1	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN same	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN cu-htk1-h1_c1	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN same	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN same	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN cu-htk1-h1_c1	MP cu-htk1-h1_c1 SI cu-htk1-h1_c1 WI cu-htk1-h1_c1 MN cu-htk1-h1_c1	MP cu-htk1-h1_c1 SI same WI cu-htk1-h1_c1 MN same
dragon1-h1_c1		MP ibm1-h1_c1 SI ibm1-h1_c1 WI ibm1-h1_c1 MN ibm1-h1_c1	MP dragon1-h1_c1 SI dragon1-h1_c1 WI dragon1-h1_c1 MN dragon1-h1_c1	MP limsi1-h1_c1 SI same WI same MN same	MP dragon1-h1_c1 SI dragon1-h1_c1 WI dragon1-h1_c1 MN dragon1-h1_c1	MP same SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP sri1-h1_c1 SI sri1-h1_c1 WI sri1-h1_c1 MN same
ibm1-h1_c1			MP ibm1-h1_c1 SI ibm1-h1_c1 WI ibm1-h1_c1 MN ibm1-h1_c1	MP ibm1-h1_c1 SI same WI same MN same	MP ibm1-h1_c1 SI ibm1-h1_c1 WI ibm1-h1_c1 MN ibm1-h1_c1	MP ibm1-h1_c1 SI ibm1-h1_c1 WI ibm1-h1_c1 MN ibm1-h1_c1	MP ibm1-h1_c1 SI ibm1-h1_c1 WI ibm1-h1_c1 MN ibm1-h1_c1	MP ibm1-h1_c1 SI ibm1-h1_c1 WI ibm1-h1_c1 MN ibm1-h1_c1	MP ibm1-h1_c1 SI ibm1-h1_c1 WI ibm1-h1_c1 MN ibm1-h1_c1	MP ibm1-h1_c1 SI same WI ibm1-h1_c1 MN same
ku1-h1_c1				MP limsi1-h1_c1 SI limsi1-h1_c1 WI limsi1-h1_c1 MN limsi1-h1_c1	MP mit-ll1-h1_c1 SI mit-ll1-h1_c1 WI mit-ll1-h1_c1 MN same	MP nyu1-h1_c1 SI nyu1-h1_c1 WI nyu1-h1_c1 MN nyu1-h1_c1	MP nyu2-h1_c1 SI nyu2-h1_c1 WI nyu2-h1_c1 MN nyu2-h1_c1	MP phil-th1-h1_c1 SI phil-th1-h1_c1 WI phil-th1-h1_c1 MN phil-th1-h1_c1	MP phil-th2-h1_c1 SI phil-th2-h1_c1 WI phil-th2-h1_c1 MN phil-th2-h1_c1	MP sri1-h1_c1 SI sri1-h1_c1 WI sri1-h1_c1 MN sri1-h1_c1
limsi1-h1_c1					MP limsi1-h1_c1 SI limsi1-h1_c1 WI limsi1-h1_c1 MN limsi1-h1_c1	MP limsi1-h1_c1 SI same WI same MN same	MP limsi1-h1_c1 SI same WI same MN same	MP limsi1-h1_c1 SI limsi1-h1_c1 WI limsi1-h1_c1 MN limsi1-h1_c1	MP limsi1-h1_c1 SI limsi1-h1_c1 WI limsi1-h1_c1 MN limsi1-h1_c1	MP same SI same WI same MN same
mit-ll1-h1_c1						MP nyu1-h1_c1 SI nyu1-h1_c1 WI nyu1-h1_c1 MN nyu1-h1_c1	MP nyu2-h1_c1 SI nyu2-h1_c1 WI nyu2-h1_c1 MN nyu2-h1_c1	MP phil-th1-h1_c1 SI phil-th1-h1_c1 WI phil-th1-h1_c1 MN phil-th1-h1_c1	MP phil-th2-h1_c1 SI phil-th2-h1_c1 WI phil-th2-h1_c1 MN phil-th2-h1_c1	MP sri1-h1_c1 SI sri1-h1_c1 WI sri1-h1_c1 MN sri1-h1_c1
nyu1-h1_c1							MP same SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP sri1-h1_c1 SI same WI same MN same
nyu2-h1_c1								MP same SI same WI same MN same	MP same SI same WI same MN same	MP sri1-h1_c1 SI same WI same MN same
phil-th1-h1_c1									MP same SI same WI same MN same	MP sri1-h1_c1 SI sri1-h1_c1 WI sri1-h1_c1 MN sri1-h1_c1
phil-th2-h1_c1										MP sri1-h1_c1 SI sri1-h1_c1 WI sri1-h1_c1 MN sri1-h1_c1
sri1-h1_c1										

Table A2 Significance Test Results: Hub 1 (20K Baseline) C1 Systems

Nov 93 Hub and Spoke CSR Evaluation
Hub 2: Telephone NAB News

GOAL: Demonstrate SI performance on unlimited-vocabulary read speech over long-distance telephone lines

DATA: 20 B speakers * 15 utts = 300 utts
- Texts drawn from same source as H1
- Digital T1 service (8-bit mu-law) at the collection end, unconstrained electronics and microphone at the subject end.

Primary and Contrast Conditions

P0 (req) any grammar or acoustic training predating June 16, 1994

	Primary P0
System	Word Err. (%)
cmu5	23.5
limsi2	24.6
sri4	22.5

Table A3 Hub 2 Results

Nov 94 Hub and Spoke CSR Evaluation
Spoke 2: Domain Adaptation

GOAL: Evaluate techniques to adapt to new news topics *not* found in training

DATA: 20 A speakers * 10 utts * 2 topics = ~400 utts, topics are outside of North American Business (NAB) News, Sennheiser mic.

Primary and Contrast Conditions

P0 (req) any grammar or acoustic training, predating June 16, 1994, but not including any training from test topics.

C1 (req) H1-P0 System on S2 Data.

C2 (req) Same as P0 with additional ~10K words of *topic-specific training.*

	Primary P0	Contrast C1	Contrast C2
System	Word Err. (%)	Word Err. (%)	Word Err. (%)
cmu3 China	17.3	17.5	*
cmu3 Simpson	21.5	22.0	*

* (Results not received at NIST as of 01/17/95)

COMPARISONS AND SIGNIFICANCE TESTS

	Test Comp.	% Reduct. W.E.	Significance Tests: McN	MAPSSWE	Sign	Wilcoxon
cmu3 China	P0:C1	1.1%	same	same	same	same
cmu3 Simpson	P0:C1	2.1%	same	same	same	same

Table A4 Spoke 2: Domain Adaptation Results

Nov 94 Hub and Spoke CSR Evaluation
Spoke 3: Non-Native Speakers

GOAL: evaluate supervised speaker adaptation on non-native speakers.

DATA: 10 C spkrs * 20 utts = 200 utts (test)

10 C spkrs * 40 utts = 400 utts common rapid enrollment data

10 C spkrs * 160 utts = 1600 utts additional rapid enrollment data

Primary and Contrast Conditions

P0	(req) speaker adaptation enabled using 40 utts rapid enrollment
C1	(req) S3_P0 system with speaker adaptation disabled
C2	(req) S3_C1 system on S0 data (native-speaker comparison)
C3	(opt) S3_P0 system on S0 data (native-speaker comparison)
C4	(opt) adaptation enabled using 100 utts rapid enrollment
C5	(opt) adaptation enabled using 200 utts rapid enrollment

SIDE INFO: session boundaries and utterance order are not known

speaker identity is known for P0, C3, C4 and C5.

	Primary P0	Contrast C1	Contrast C2	Contrast C3	Contrast C4	Contrast C5
System	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)	Word Err. (%)
bbn2	10.1 *	26.1 *	8.1 *	6.4 *		
cu-htk2	11.0	20.7	5.7			
sri2	11.1	22.9	7.9		9.2	7.8

* (Late: Re-submitted using correct LM)

COMPARISONS AND SIGNIFICANCE TESTS

	Test Comp.	% Reduct. W.E.	McN	MAPSSWE	Sign	Wilcoxon
			Significance Tests:			
bbn2	P0:C1	61.3%	P0	P0	P0	P0
cu-htk2	P0:C1	47.0%	P0	P0	P0	P0
sri2	P0:C1	51.7%	P0	P0	P0	P0

	Test Comp.	% Reduct. W.E.	McN	MAPSSWE	Sign	Wilcoxon
			Significance Tests:			
sri2	C4:P0	17.2%	C4	C4	C4	C4

	Test Comp.	% Reduct. W.E.	McN	MAPSSWE	Sign	Wilcoxon
			Significance Tests:			
sri2	C5:C4	14.9%	C5	C5	same	C5

	Test Comp.	% Reduct. W.E.	McN	MAPSSWE	Sign	Wilcoxon
			Significance Tests:			
sri2	C5:P0	29.6%	C5	C5	C5	C5

Table A5 Spoke 3: Recognition Outlier Results

Nov 94 Hub and Spoke CSR Evaluation
Spoke 4: Incremental Speaker Adaptation

GOAL: evaluate an incremental speaker adaptation algorithm.

DATA: 4 A spkrs * 100 utts = 400 utts

5K-word read WSJ data, Sennheiser mic.

Primary and Contrast Conditions

P0 (req) incremental unsupervised speaker adaptation

C1 (req) S4-P0 system with speaker adaptation disabled

C2 (opt) supervised incremental adaptation

SIDE INFO: correct transcript is known after the fact for C2.

	Primary P0	Contrast C1	Contrast C2
System	Word Err. (%)	Word Err. (%)	Word Err. (%)
bbn3 Utts 1-25	11.8 *	11.9 *	11.3 *
bbn3 Utts 26-50	10.3 *	11.7 *	10.2 *
bbn3 Utts 51-75	9.0 *	12.0 *	9.0 *
bbn3 Utts 76+	9.8 *	11.5 *	9.1 *
cu-htk3 Utts 1-25	7.2 @	8.2	
cu-htk3 Utts 26-50	7.8 @	8.4	
cu-htk3 Utts 51-75	5.8 @	7.2	
cu-htk3 Utts 76+	5.0 @	7.1	
dragon2 Utts 1-25	10.6	10.9	10.1
dragon2 Utts 26-50	11.1	12.4	10.8
dragon2 Utts 51-75	9.0	10.1	8.7
dragon2 Utts 76+	10.2	11.4	9.2

* (Late: Re-submitted using correct LM)
@ (Late: Re-submitted after hardware repair)

COMPARISONS AND SIGNIFICANCE TESTS

	Test	% Reduct.	Significance Tests:	
	Comp.	W.E.	McN	MAPSSWE
bbn3 Utts 1-25	P0:C1	0.5%	same	same
bbn3 Utts 26-50	P0:C1	11.6%	same	P0
bbn3 Utts 51-75	P0:C1	25.0%	P0	P0
bbn3 Utts 76+	P0:C1	15.3%	same	same
cu-htk3 Utts 1-25	P0:C1	12.2%	same	P0
cu-htk3 Utts 26-50	P0:C1	6.7%	same	same
cu-htk3 Utts 51-75	P0:C1	19.4%	same	P0
cu-htk3 Utts 76+	P0:C1	30.4%	P0	P0
dragon2 Utts 1-25	P0:C1	2.3%	same	same
dragon2 Utts 26-50	P0:C1	10.5%	same	P0
dragon2 Utts 51-75	P0:C1	10.3%	P0	P0
dragon2 Utts 76+	P0:C1	11.0%	same	same

	Runtime Ratio s4_p0/s4_c1
bbn3	1.05
cu-htk3	1.33
dragon2	0.86

	Test	% Reduct.	Significance Tests:	
	Comp.	W.E.	McN	MAPSSWE
bbn3 Utts 1-25	C2:P0	4.2%	same	same
bbn3 Utts 26-50	C2:P0	1.1%	same	same
bbn3 Utts 51-75	C2:P0	0.6%	same	same
bbn3 Utts 76+	C2:P0	6.4%	same	C2
dragon2 Utts 1-25	C2:P0	4.7%	same	same
dragon2 Utts 26-50	C2:P0	3.0%	same	same
dragon2 Utts 51-75	C2:P0	3.2%	same	same
dragon2 Utts 76+	C2:P0	9.6%	same	C2

Table A6 Spoke 4: Incremental Speaker Adaptation Results

```
                 Nov 94 Hub and Spoke CSR Evaluation
                   Spoke 5: Microphone Independence

   GOAL:      evaluate an unsupervised channel compensation
              algorithm
   DATA:      20 A spkrs * 10 utts = 200 utts (2 channels, read
              speech from Spoke 0, 5K-word read WSJ data)

              10 different mics not in training or development
              test.  NOTE: No speech from the test microphones
              can be used.

                  Primary and Contrast Conditions

   P0         (req) unsupervised channel compensation enabled on
              wv2 data

   C1         (req) S5-P0 system with compensation disabled on wv2
              data

   C2         (req) S5-P0 system on Sennheiser (wv1) data

   C3         (req) S5-C1 system on Sennheiser (wv1) data

SIDE INFO:    Microphone identities are not known
-----------------------------------------------------------------
       | Primary P0   | Contrast C1  | Contrast C2  | Contrast C3
-------+--------------+--------------+--------------+------------
System | Word Err. (%)| Word Err. (%)| Word Err. (%)| Word Err. (%)
-------+--------------+--------------+--------------+------------
cmu2   |     9.9      |     12.4     |     6.5      |     6.7
cmu4   |     9.7      |     12.4     |     7.1      |     6.7
=================================================================

                COMPARISONS AND SIGNIFICANCE TESTS
-----------------------------------------------------------------
         | Test     % Reduct. |       Significance Tests:
         | Comp.      W.E.    | McN   MAPSSWE    Sign    Wilcoxon
         |                    |
cmu2     | P0:C1      20.1%   | P0     P0        same     P0
cmu4     | P0:C1      21.4%   | same   P0        P0       P0
---------+--------------------+----------------------------------
         | Test     % Reduct. |       Significance Tests:
         | Comp.      W.E.    | McN   MAPSSWE    Sign    Wilcoxon
         |                    |
cmu2     | C2:P0      34.0%   | C2     C2        C2       C2
cmu4     | C2:P0      26.8%   | C2     C2        C2       C2
---------+--------------------+----------------------------------
         | Test     % Reduct. |       Significance Tests:
         | Comp.      W.E.    | McN   MAPSSWE    Sign    Wilcoxon
         |                    |
cmu2     | C3:C2      -2.4%   | same   same      same     same
cmu4     | C3:C2       6.1%   | same   same      same     same
---------+--------------------+----------------------------------
         | Test     % Reduct. |       Significance Tests:
         | Comp.      W.E.    | McN   MAPSSWE    Sign    Wilcoxon
         |                    |
cmu2     | C3:C1      46.0%   | C3     C3        C3       C3
cmu4     | C3:C1      46.0%   | C3     C3        C3       C3
```

Table A7 Spoke 5: Microphone Independence Results

```
               Nov 94 Hub and Spoke CSR Evaluation
               Spoke 9: Spontaneous WSJ Dictation

GOAL:     improve basic SI performance on spontaneous
          dictation of business news stories.
DATA:     10 D speakers * 20 utts = 200 utts
          Spontaneous dictations of business news stories,
          Sennheiser mic

                 Primary and Contrast Conditions

 P0       (req) any grammar or acoustic training predating
          June 16, 1994

 C1       (req) H1-P0 System on S9 data
-------------------------------------------------------------
               |      Primary P0      |      Contrast C1
---------------+----------------------+----------------------
    System     |    Word Err. (%)     |    Word Err. (%)
---------------+----------------------+----------------------
    bbn4       |        14.2          |        14.2
=============================================================

              COMPARISONS AND SIGNIFICANCE TESTS
-------------------------------------------------------------
       | Test    % Reduct. |       Significance Tests:
       | Comp.     W.E.    | McN   MAPSSWE   Sign   Wilcoxon
       |                   |
bbn4   | P0:C1     0.0%    | same   same     same     same
```

Table A8 Spoke 9: Spontaneous WSJ Dictation Results

Nov 94 Hub and Spoke CSR Evaluation
Spoke 10: Noisy Channel

GOAL: Evaluate compensation on data corrupted with additive noise
DATA: 10 A speakers * 10 utts * 3 levels = 300 utts for P0 and C1
10 A Speakers * 10 utts = 100 utts for C2, Sennheiser mic.

Primary and Contrast Conditions

P0 (req) noise compensation enabled for 3 SNR levels

C1 (req) S10-P0 system with compensation disabled for 3 SNR levels

C2 (req) S10-C1 system on 'Clean' data

	Primary 0			Contrast 1			Contrast C2
	22DB	16DB	10DB	22DB	16DB	10DB	
System	W.E. (%)	W.E. (%)	W.E. (%)	W.E. (%)	W.E. (%)	W.E. (%)	W.E. (%)
cu-htk4	9.4	13.4	19.8	41.9	59.4	84.7	7.2
ibm2	8.4	10.0	12.8	15.4	42.2	77.4	7.2
sri3	8.4	9.8	12.2	11.1	18.4	35.4	6.7

COMPARISONS AND SIGNIFICANCE TESTS

	Test Comp.	% Reduct. W.E.	McN	Significance Tests: MAPSSWE	Sign	Wilcoxon
cu-htk4	P0_22DB:C1_22DB	77.6%	P0_22DB	P0_22DB	P0_22DB	P0_22DB
ibm2	P0_22DB:C1_22DB	45.6%	P0_22DB	P0_22DB	P0_22DB	P0_22DB
sri3	P0_22DB:C1_22DB	24.1%	P0_22DB	P0_22DB	same	P0_22DB
	Test Comp.	% Reduct. W.E.	McN	Significance Tests: MAPSSWE	Sign	Wilcoxon
cu-htk4	P0_16DB:C1_16DB	77.5%	P0_16DB	P0_16DB	P0_16DB	P0_16DB
ibm2	P0_16DB:C1_16DB	76.2%	P0_16DB	P0_16DB	P0_16DB	P0_16DB
sri3	P0_16DB:C1_16DB	46.7%	P0_16DB	P0_16DB	P0_16DB	P0_16DB
	Test Comp.	% Reduct. W.E.	McN	Significance Tests: MAPSSWE	Sign	Wilcoxon
cu-htk4	P0_10DB:C1_10DB	76.6%	P0_10DB	P0_10DB	P0_10DB	P0_10DB
ibm2	P0_10DB:C1_10DB	83.5%	P0_10DB	P0_10DB	P0_10DB	P0_10DB
sri3	P0_10DB:C1_10DB	65.4%	P0_10DB	P0_10DB	P0_10DB	P0_10DB
	Test Comp.	% Reduct. W.E.	McN	Significance Tests: MAPSSWE	Sign	Wilcoxon
cu-htk4	C2:P0_22DB	23.4%	same	C2	same	C2
ibm2	C2:P0_22DB	13.7%	same	same	same	same
sri3	C2:P0_22DB	20.1%	same	C2	C2	C2
	Test Comp.	% Reduct. W.E.	McN	Significance Tests: MAPSSWE	Sign	Wilcoxon
cu-htk4	C2:P0_16DB	46.3%	C2	C2	C2	C2
ibm2	C2:P0_16DB	27.9%	C2	C2	same	C2
sri3	C2:P0_16DB	31.3%	C2	C2	C2	C2
	Test Comp.	% Reduct. W.E.	McN	Significance Tests: MAPSSWE	Sign	Wilcoxon
cu-htk4	C2:P0_10DB	63.7%	C2	C2	C2	C2
ibm2	C2:P0_10DB	43.3%	C2	C2	C2	C2
sri3	C2:P0_10DB	44.8%	C2	C2	C2	C2
	Test Comp.	% Reduct. W.E.	McN	Significance Tests: MAPSSWE	Sign	Wilcoxon
cu-htk4	C2:C1_22DB	82.9%	C2	C2	C2	C2
ibm2	C2:C1_22DB	53.0%	C2	C2	C2	C2
sri3	C2:C1_22DB	39.4%	C2	C2	C2	C2
	Test Comp.	% Reduct. W.E.	McN	Significance Tests: MAPSSWE	Sign	Wilcoxon
cu-htk4	C2:C1_16DB	87.9%	C2	C2	C2	C2
ibm2	C2:C1_16DB	82.9%	C2	C2	C2	C2
sri3	C2:C1_16DB	63.4%	C2	C2	C2	C2
	Test Comp.	% Reduct. W.E.	McN	Significance Tests: MAPSSWE	Sign	Wilcoxon
cu-htk4	C2:C1_10DB	91.5%	C2	C2	C2	C2
ibm2	C2:C1_10DB	90.6%	C2	C2	C2	C2
sri3	C2:C1_10DB	80.9%	C2	C2	C2	C2

Table A9 Spoke 10: Noisy Channel Results

Dec93 ATIS SPREC Test Results

Class A+D+X Subset

	W. Err	Corr	Sub	Del	Ins	U. Err	# Utt.
att2-adx	3.5	97.2	1.9	0.9	0.7	21.4	981
bbn3-adx	3.5	97.4	1.9	0.7	0.9	20.7	981
cmu3-adx	3.4	97.3	1.9	0.8	0.7	17.7	981
cmu4-adx *	3.4	97.3	2.0	0.7	0.7	18.1	981
mit_lcs3-adx	5.2	95.4	2.5	2.0	0.6	27.6	981
mit_lcs4-adx *	5.1	95.5	2.5	2.0	0.6	27.8	981
mitre2-adx	14.8	86.5	7.6	5.9	1.2	56.3	981
sri3-adx	2.5	98.3	1.3	0.5	0.8	15.5	981
sri4-adx *	2.1	98.5	1.1	0.4	0.6	13.1	981
unisys2-adx	4.1	97.0	2.4	0.6	1.1	26.1	981

Class A+D Subset

	W. Err	Corr	Sub	Del	Ins	U. Err	# Utt.
att2-a_d	3.0	97.6	1.5	0.9	0.6	19.0	732
bbn3-a_d	3.0	97.7	1.6	0.7	0.8	18.9	732
cmu3-a_d	3.3	97.3	1.8	0.9	0.6	16.8	732
cmu4-a_d *	3.1	97.5	1.7	0.8	0.6	16.8	732
mit_lcs3-a_d	4.4	96.2	2.2	1.6	0.6	25.0	732
mit_lcs4-a_d *	4.4	96.2	2.2	1.6	0.6	25.1	732
mitre2-a_d	14.1	87.1	6.9	6.0	1.2	54.8	732
sri3-a_d	2.3	98.4	1.1	0.5	0.7	15.2	732
sri4-a_d *	1.9	98.6	0.9	0.5	0.6	12.8	732
unisys2-a_d	3.6	97.4	2.0	0.7	1.0	22.7	732

Class A Subset

	W. Err	Corr	Sub	Del	Ins	U. Err	# Utt.
att2-a	2.6	97.9	1.4	0.7	0.5	18.0	445
bbn3-a	2.8	98.0	1.5	0.5	0.8	19.1	445
cmu3-a	3.0	97.6	1.6	0.8	0.6	17.5	445
cmu4-a *	2.7	97.8	1.5	0.6	0.6	16.6	445
mit_lcs3-a	4.1	96.5	2.0	1.5	0.6	24.9	445
mit_lcs4-a *	4.0	96.6	1.9	1.5	0.6	24.9	445
mitre2-a	13.4	88.0	6.8	5.1	1.4	56.6	445
sri3-a	2.3	98.5	1.1	0.4	0.7	16.0	445
sri4-a *	1.9	98.7	0.9	0.4	0.6	13.7	445
unisys2-a	3.1	97.8	1.7	0.5	0.9	20.7	445

Class D Subset

	W. Err	Corr	Sub	Del	Ins	U. Err	# Utt.
att2-d	3.7	97.0	1.8	1.3	0.7	20.6	287
bbn3-d	3.5	97.1	1.7	1.1	0.6	18.5	287
cmu3-d	3.8	96.7	2.0	1.3	0.5	15.7	287
cmu4-d *	3.7	96.9	2.0	1.1	0.7	17.1	287
mit_lcs3-d	5.3	95.4	2.6	2.0	0.7	25.1	287
mit_lcs4-d *	5.3	95.4	2.8	1.8	0.7	25.4	287
mitre2-d	15.5	85.1	7.1	7.8	0.6	51.9	287
sri3-d	2.2	98.4	1.0	0.7	0.6	13.9	287
sri4-d *	2.0	98.5	0.9	0.6	0.5	11.5	287
unisys2-d	4.7	96.5	2.5	1.0	1.2	25.8	287

Class X Subset

	W. Err	Corr	Sub	Del	Ins	U. Err	# Utt.
att2-x	4.9	96.0	3.0	1.0	0.9	28.5	249
bbn3-x	4.7	96.5	2.9	0.6	1.1	26.1	249
cmu3-x	3.9	97.1	2.4	0.5	1.0	20.5	249
cmu4-x *	4.2	96.8	2.7	0.5	1.0	22.1	249
mit_lcs3-x	7.3	93.4	3.5	3.1	0.7	35.3	249
mit_lcs4-x *	7.3	93.4	3.5	3.1	0.7	35.7	249
mitre2-x	16.8	84.7	9.7	5.6	1.5	60.6	249
sri3-x	3.2	97.8	1.9	0.3	1.0	16.5	249
sri4-x *	2.6	98.1	1.7	0.2	0.7	14.1	249
unisys2-x	5.7	95.8	3.7	0.5	1.5	36.1	249

Note: All systems are "Primary" unless otherwise indicated

* "Comparative Systems"

Table A10 ATIS SPREC Benchmark Test Results

Dec93 ATIS SPREC Test Results

		Class A+D Subset Originating Site of Test Data							
		BBN (137 Utt.)	CMU (128 Utt.)	MIT (165 Utt.)	NIST-BBN (75 Utt.)	NIST-SRI (86 Utt.)	SRI (141 Utt.)	Overall Totals 732	Foreign Coll. Site Totals
	att2	1.3 0.4 0.4 2.0 17.5	1.3 1.2 0.3 2.8 20.3	1.3 0.3 1.1 2.6 17.0	1.1 0.3 0.2 1.5 8.0	1.7 0.8 0.7 3.2 18.6	2.3 2.4 0.4 5.1 27.7	1.5 0.9 0.6 3.0 19.0	1.5 0.9 0.6 3.0 19.0
	bbn3	1.0 0.2 0.5 1.7 13.1	1.2 0.7 0.5 2.4 16.4	1.3 0.1 1.3 2.6 20.6	0.3 0.2 0.8 1.2 8.0	1.9 0.8 0.5 3.2 22.1	3.3 2.3 0.7 6.3 28.4	1.6 0.7 0.8 3.0 18.9	1.7 0.8 0.8 3.4 20.2
	cmu3	1.1 0.4 0.5 2.0 13.1	0.9 0.4 0.5 1.8 12.5	1.2 0.4 0.9 2.5 17.6	0.3 0.3 0.2 0.8 5.3	2.0 0.7 0.8 3.5 19.8	4.4 3.2 0.3 8.0 27.7	1.8 0.9 0.6 3.3 16.8	1.9 1.0 0.6 3.6 17.7
	cmu4	0.9 0.2 0.5 1.6 12.4	1.0 0.5 0.6 2.0 13.3	1.0 0.3 0.7 2.0 15.8	0.8 0.2 0.0 0.9 6.7	2.4 0.7 1.3 4.4 23.3	4.2 2.5 0.5 7.2 27.0	1.7 0.8 0.6 3.1 16.8	1.8 0.8 0.6 3.3 17.5
SYSTEMS	mit_lcs3	1.9 0.6 0.6 3.1 17.5	1.8 2.0 0.3 4.1 28.1	1.6 0.9 1.0 3.5 25.5	0.9 1.7 0.2 2.7 12.0	1.7 1.1 1.1 3.9 20.9	4.5 3.6 0.3 8.5 38.3	2.2 1.6 0.6 4.4 25.0	2.4 1.9 0.5 4.8 24.9
	mit_lcs4	2.0 0.6 0.4 3.1 16.8	2.0 2.4 0.3 4.8 32.8	1.7 0.8 1.1 3.5 26.7	1.1 1.7 0.3 3.1 12.0	1.7 0.7 0.9 3.4 18.6	4.1 3.2 0.3 7.6 35.5	2.2 1.6 0.6 4.4 25.1	2.4 1.8 0.4 4.7 24.7
	mitre2	4.1 2.2 0.6 6.9 40.9	4.8 3.3 1.6 9.7 47.7	6.4 4.0 1.6 12.0 61.8	4.0 3.1 0.6 7.6 26.7	6.0 2.1 2.0 10.2 44.2	14.2 18.4 0.5 33.1 87.9	6.9 6.0 1.2 14.1 54.8	6.9 6.0 1.2 14.1 54.8
	sri3	0.4 0.0 0.6 1.0 7.3	0.6 0.5 0.8 1.9 17.2	1.2 0.3 0.9 2.3 19.4	0.2 0.0 0.5 0.6 5.3	0.9 0.5 0.5 2.0 12.8	2.5 1.6 0.5 4.6 22.7	1.1 0.5 0.7 2.3 15.2	0.7 0.2 0.7 1.7 13.4
	sri4	0.3 0.0 0.5 0.8 5.8	0.8 0.5 0.6 1.9 16.4	0.9 0.2 0.9 1.9 15.2	0.2 0.0 0.5 0.6 5.3	0.9 0.5 0.5 2.0 12.8	1.9 1.5 0.3 3.7 17.7	0.9 0.5 0.6 1.9 12.8	0.6 0.2 0.6 1.5 11.7
	unisys2	1.3 0.1 0.8 2.3 17.5	1.3 0.6 0.7 2.7 19.5	1.9 0.2 1.5 3.5 26.7	0.5 0.2 0.6 1.2 9.3	2.4 0.8 1.2 4.4 29.1	3.6 2.2 0.7 6.5 29.1	2.0 0.7 1.0 3.6 22.7	2.0 0.7 1.0 3.6 22.7
Overall Totals		1.4 0.5 0.5 2.4 16.2	1.6 1.2 0.6 3.4 22.4	1.8 0.7 1.1 3.6 24.6	0.9 0.7 0.4 2.0 9.9	2.2 0.9 1.0 4.0 22.2	4.5 4.1 0.5 9.1 34.2		
Foreign System		1.5 0.5 0.5 2.5 16.5	1.7 1.4 0.6 3.8 24.8	1.9 0.7 1.1 3.7 24.2	0.9 0.7 0.4 2.0 9.9	2.2 0.9 1.0 4.0 22.2	5.1 4.7 0.5 10.3 37.7		%Sub %Del %Ins %W.Err %Utt.Err

Matrix tabulation of results for the *Dec93 ATIS SPREC Test Results*, for the Class A+D Subset.

Matrix columns present results for Test Data Subsets collected at several sites, and matrix rows present results for different systems.

Numbers printed at the top of the matrix columns indicate the number of utterances in the Test Data (sub)set from the corresponding site.

"Overall Totals" (column) present results for the entire Class A+D Subset for the system corresponding to that matrix row. "Foreign Coll. Site Totals" present results for "foreign site" data (i.e., excluding locally collected data) for the Class A+D Subset.

"Overall Totals" (row) present results accumulated over all systems corresponding to the Test Data (sub)set corresponding to that matrix column. "Foreign System Totals" present results accumulated over "foreign systems" (i.e., excluding results for the system(s) developed at the site responsible for collection of that Test Data subset.)

Table A11 ATIS SPREC Results: Class (A+D) by Collection Site

Composite Report of All Significance Tests
For the Dec93 ATIS SPREC Class A+D Test Results Test

Test Name	Abbrev.
Matched Pair Sentence Segment (Word Error) Test	MP
Signed Paired Comparison (Speaker Word Accuracy Rate (%)) Test	SI
Wilcoxon Signed Rank (Speaker Word Accuracy Rate (%)) Test	WI
McNemar (Sentence Error) Test	MN

	att2-a_d	bbn3-a_d	cmu3-a_d	cmu4-a_d	mit_lcs3-a_d	mit_lcs4-a_d	mitre2-a_d	sri3-a_d	sri4-a_d	unisys2-a_d
att2-a_d		MP same SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP att2-a_d SI att2-a_d WI att2-a_d MN att2-a_d	MP att2-a_d SI att2-a_d WI att2-a_d MN att2-a_d	MP att2-a_d SI att2-a_d WI att2-a_d MN att2-a_d	MP sri3-a_d SI same WI sri3-a_d MN sri3-a_d	MP sri4-a_d SI sri4-a_d WI sri4-a_d MN sri4-a_d	MP same SI same WI same MN att2-a_d
bbn3-a_d			MP same SI same WI same MN same	MP same SI cmu4-a_d WI same MN same	MP bbn3-a_d SI bbn3-a_d WI bbn3-a_d MN bbn3-a_d	MP bbn3-a_d SI bbn3-a_d WI bbn3-a_d MN bbn3-a_d	MP bbn3-a_d SI bbn3-a_d WI bbn3-a_d MN bbn3-a_d	MP sri3-a_d SI same WI same MN sri3-a_d	MP sri4-a_d SI sri4-a_d WI sri4-a_d MN sri4-a_d	MP bbn3-a_d SI bbn3-a_d WI bbn3-a_d MN bbn3-a_d
cmu3-a_d				MP same SI same WI same MN same	MP cmu3-a_d SI cmu3-a_d WI cmu3-a_d MN cmu3-a_d	MP cmu3-a_d SI cmu3-a_d WI cmu3-a_d MN cmu3-a_d	MP cmu3-a_d SI cmu3-a_d WI cmu3-a_d MN cmu3-a_d	MP sri3-a_d SI same WI same MN same	MP sri4-a_d SI sri4-a_d WI sri4-a_d MN sri4-a_d	MP same SI cmu3-a_d WI cmu3-a_d MN cmu3-a_d
cmu4-a_d					MP cmu4-a_d SI cmu4-a_d WI cmu4-a_d MN cmu4-a_d	MP cmu4-a_d SI same WI cmu4-a_d MN cmu4-a_d	MP cmu4-a_d SI cmu4-a_d WI cmu4-a_d MN cmu4-a_d	MP sri3-a_d SI same WI same MN same	MP sri4-a_d SI same WI same MN sri4-a_d	MP same SI cmu4-a_d WI cmu4-a_d MN cmu4-a_d
mit_lcs3-a_d						MP same SI same WI same MN same	MP mit_lcs3-a_d SI mit_lcs3-a_d WI mit_lcs3-a_d MN mit_lcs3-a_d	MP sri3-a_d SI sri3-a_d WI sri3-a_d MN sri3-a_d	MP sri4-a_d SI sri4-a_d WI sri4-a_d MN sri4-a_d	MP unisys2-a_d SI unisys2-a_d WI same MN same
mit_lcs4-a_d							MP mit_lcs4-a_d SI mit_lcs4-a_d WI mit_lcs4-a_d MN mit_lcs4-a_d	MP sri3-a_d SI sri3-a_d WI sri3-a_d MN sri3-a_d	MP sri4-a_d SI sri4-a_d WI sri4-a_d MN sri4-a_d	MP unisys2-a_d SI unisys2-a_d WI same MN same
mitre2-a_d								MP sri3-a_d SI sri3-a_d WI sri3-a_d MN sri3-a_d	MP sri4-a_d SI sri4-a_d WI sri4-a_d MN sri4-a_d	MP unisys2-a_d SI unisys2-a_d WI unisys2-a_d MN unisys2-a_d
sri3-a_d									MP sri4-a_d SI sri4-a_d WI sri4-a_d MN sri4-a_d	MP sri3-a_d SI same WI sri3-a_d MN sri3-a_d
sri4-a_d										MP sri4-a_d SI sri4-a_d WI sri4-a_d MN sri4-a_d
unisys2-a_d										

Table A12 Significance Test Results: ATIS SPREC Class (A+D) Systems

Dec 94 ATIS NL Test Results - Using Minimal/Maximal Scoring Criterion

Class A+D

system	UW. Err	# T	# F	# NA	# Utt
att1	5.9	689	43	0	732
bbn1	13.9	630	102	0	732
cmu1	6.3	686	46	0	732
mit_lcs1	7.1	680	52	0	732
mitre1	41.7	427	305	0	732
sri1	10.7	654	70	8	732
unisys1	33.5	487	145	100	732

Class A

system	UW. Err	# T	# F	# NA	# Utt
att1-a	3.8	428	17	0	445
bbn1-a	9.4	403	42	0	445
cmu1-a	3.8	428	17	0	445
mit_lcs1-a	4.5	425	20	0	445
mitre1-a	30.6	309	136	0	445
sri1-a	7.0	414	27	4	445
unisys1-a	23.6	340	65	40	445

Class D

system	UW. Err	# T	# F	# NA	# Utt
att1-d	9.1	261	26	0	287
bbn1-d	20.9	227	60	0	287
cmu1-d	10.1	258	29	0	287
mit_lcs1-d	11.1	255	32	0	287
mitre1-d	58.9	118	169	0	287
sri1-d	16.4	240	43	4	287
unisys1-d	48.8	147	80	60	287

Table A13 ATIS NL Results

Dec 94 ATIS NL Test Results - Using Minimal/Maximal Scoring Criterion

		Class (A+D) Set Originating Site of Test Data							
		BBN 137	CMU 128	MIT 165	NIST-BBN 75	NIST-SRI 86	SRI 141	Overall Totals 732	Foreign Coll. Site Totals
SYSTEMS	att1	127 10 0 93 7 0 7.3	125 3 0 98 2 0 2.3	150 15 0 91 9 0 9.1	72 3 0 96 4 0 4.0	80 6 0 93 7 0 7.0	135 6 0 96 4 0 4.3	689 43 0 94 6 0 5.9	689 43 0 94 6 0 5.9
	bbn1	113 24 0 82 18 0 17.5	105 23 0 82 18 0 18.0	136 29 0 82 18 0 17.6	69 6 0 92 8 0 8.0	75 11 0 87 13 0 12.8	132 9 0 94 6 0 6.4	630 102 0 86 14 0 13.9	517 78 0 87 13 0 13.1
	cmu1	128 9 0 93 7 0 6.6	122 6 0 95 5 0 4.7	149 16 0 90 10 0 9.7	72 3 0 96 4 0 4.0	83 3 0 97 3 0 3.5	132 9 0 94 6 0 6.4	686 46 0 94 6 0 6.3	564 40 0 93 7 0 6.6
	mit_lcs1	120 17 0 88 12 0 12.4	116 12 0 91 9 0 9.4	151 14 0 92 8 0 8.5	75 0 0 100 0 0 0.0	81 5 0 94 6 0 5.8	137 4 0 97 3 0 2.8	680 52 0 93 7 0 7.1	529 38 0 93 7 0 6.7
	mitre1	76 61 0 55 45 0 44.5	66 62 0 52 48 0 48.4	99 66 0 60 40 0 40.0	55 20 0 73 27 0 26.7	45 41 0 52 48 0 47.7	86 55 0 61 39 0 39.0	427 305 0 58 42 0 41.7	427 305 0 58 42 0 41.7
	sri1	125 12 0 91 9 0 8.8	113 13 2 88 10 2 11.7	143 21 1 87 13 1 13.3	73 2 0 97 3 0 2.7	73 8 5 85 9 6 15.1	127 14 0 90 10 0 9.9	654 70 8 89 10 1 10.7	527 56 8 89 9 1 10.8
	unisys1	92 19 26 67 14 19 32.8	89 22 17 70 17 13 30.5	101 41 23 61 25 14 38.8	61 9 5 81 12 7 18.7	56 19 11 65 22 13 34.9	88 35 18 62 25 13 37.6	487 145 100 67 20 14 33.5	487 145 100 67 20 14 33.5
Overall Totals		781 152 26 81 16 3 18.6	736 141 19 82 16 2 17.9	929 202 24 80 17 2 19.6	477 43 5 91 8 1 9.1	493 93 16 82 15 3 18.1	837 132 18 85 13 2 15.2		
Foreign System Totals		668 128 26 81 16 3 18.7	614 135 19 80 18 2 20.1	778 188 24 79 19 2 21.4	477 43 5 91 8 1 9.1	493 93 16 82 15 3 18.1	710 118 18 84 14 2 16.1		

Legend:

```
#T        #F  #NA
%T        %F  %NA
% Un-Weighted Err
```

Matrix tabulation of results for the Dec 94 ATIS NL Test Results - Using Minimal/Maximal Scoring Criterion, for the Class (A+D) Subset.

Matrix columns present results for Test Data Subsets collected at several sites, and matrix rows present results for different systems.

Numbers printed at the top of the matrix columns indicate the number of evaluable utterances in the Test Data (sub)set from the corresponding site.

"Overall Totals" (column) present results for the entire Class (A+D) Subset for the system corresponding to that matrix row. "Foreign Coll. Site Totals" present results for "foreign site" data (i.e., excluding locally collected data) for the Class (A+D) Subset.

"Overall Totals" (row) present results accumulated over all systems corresponding to the Test Data (sub)set corresponding to that matrix column. "Foreign System Totals" present results accumulated over "foreign systems" (i.e., excluding results for the system(s) developed at the site responsible for collection of that Test Data subset.)

Table A14 ATIS NL Results: Class (A+D) by Collection Site

COMPARISON MATRIX: McNEMAR'S TEST ON CORRECTLY ANSWERED UTTERANCES FOR THE TEST:
Dec 94 ATIS NL Test Results - Using Minimal/Maximal Scoring Criterion
Class A+D

	att1	bbn1	cmu1	mit_lcs1	mitre1	sri1	unisys1
att1		att1	same	same	att1	att1	att1
bbn1			cmu1	mit_lcs1	bbn1	sri1	bbn1
cmu1				same	cmu1	cmu1	cmu1
mit_lcs1					mit_lcs1	mit_lcs1	mit_lcs1
mitre1						sri1	unisys1
sri1							sri1
unisys1							

Table A15 ATIS NL Class (A+D) McNemar Significance Test Comparisons

Dec 94 ATIS SLS Test Results - Using Minimal/Maximal Scoring Criterion

Class A+D

system	UW. Err	# T	# F	# NA	# Utt
att1	8.9	667	65	0	732
bbn2	16.5	611	121	0	732
cmu1	9.7	661	71	0	732
cmu2 *	8.6	669	63	0	732
mit_lcs2	13.1	636	96	0	732
mitre1	55.3	327	405	0	732
sri1	13.7	632	92	8	732
sri2 *	12.7	639	86	7	732
unisys1	36.3	466	185	81	732

Class A

system	UW. Err	# T	# F	# NA	# Utt
att1-a	7.0	414	31	0	445
bbn2-a	11.9	392	53	0	445
cmu1-a	7.4	412	33	0	445
cmu2-a *	6.5	416	29	0	445
mit_lcs2-a	10.3	399	46	0	445
mitre1-a	44.9	245	200	0	445
sri1-a	10.6	398	44	3	445
unisys1-a	27.4	323	84	38	445
sri2-a *	9.7	402	40	3	445

Class D

system	UW. Err	# T	# F	# NA	# Utt
att1-d	11.8	253	34	0	287
bbn2-d	23.7	219	68	0	287
cmu1-d	13.2	249	38	0	287
cmu2-d *	11.8	253	34	0	287
mit_lcs2-d	17.4	237	50	0	287
mitre1-d	71.4	82	205	0	287
sri1-d	18.5	234	48	5	287
sri2-d *	17.4	237	46	4	287
unisys1-d	50.2	143	101	43	287

Note: All systems are "Primary" unless otherwise indicated

* "Comparative Systems"

Table A16 ATIS SLS Benchmark Test Results

Dec 94 ATIS SLS Test Results - Using Minimal/Maximal Scoring Criterion

		Class (A+D) Set Originating Site of Test Data							
		BBN 137	CMU 128	MIT 165	NIST-BBN 75	NIST-SRI 86	SRI 141	Overall Totals 732	Foreign Coll. Site Totals
SYSTEMS	att1	122 15 0 89 11 0 10.9	124 4 0 97 3 0 3.1	146 19 0 88 12 0 11.5	70 5 0 93 7 0 6.7	76 10 0 88 12 0 11.6	129 12 0 91 9 0 8.5	667 65 0 91 9 0 8.9	667 65 0 91 9 0 8.9
	bbn2	114 23 0 83 17 0 16.8	102 26 0 80 20 0 20.3	129 36 0 78 22 0 21.8	68 7 0 91 9 0 9.3	73 13 0 85 15 0 15.1	125 16 0 89 11 0 11.3	611 121 0 83 17 0 16.5	497 98 0 84 16 0 16.5
	cmu1	124 13 0 91 9 0 9.5	116 12 0 91 9 0 9.4	148 17 0 90 10 0 10.3	72 3 0 96 4 0 4.0	76 10 0 88 12 0 11.6	125 16 0 89 11 0 11.3	661 71 0 90 10 0 9.7	545 59 0 90 10 0 9.8
	cmu2	125 12 0 91 9 0 8.8	117 11 0 91 9 0 8.6	146 19 0 88 12 0 11.5	71 4 0 95 5 0 5.3	80 6 0 93 7 0 7.0	130 11 0 92 8 0 7.8	669 63 0 91 9 0 8.6	552 52 0 91 9 0 8.6
	mit_lcs2	117 20 0 85 15 0 14.6	109 19 0 85 15 0 14.8	138 27 0 84 16 0 16.4	71 4 0 95 5 0 5.3	79 7 0 92 8 0 8.1	122 19 0 87 13 0 13.5	636 96 0 87 13 0 13.1	498 69 0 88 12 0 12.2
	mitre1	60 77 0 44 56 0 56.2	55 73 0 43 57 0 57.0	79 86 0 48 52 0 52.1	50 25 0 67 33 0 33.3	34 52 0 40 60 0 60.5	49 92 0 35 65 0 65.2	327 405 0 45 55 0 55.3	327 405 0 45 55 0 55.3
	sri1	120 17 0 88 12 0 12.4	108 17 3 84 13 2 15.6	135 30 0 82 18 0 18.2	73 2 0 97 3 0 2.7	72 9 5 84 10 6 16.3	124 17 0 88 12 0 12.1	632 92 8 86 13 1 13.7	508 75 8 86 13 1 14.0
	sri2	123 14 0 90 10 0 10.2	109 17 2 85 13 2 14.8	138 27 0 84 16 0 16.4	73 2 0 97 3 0 2.7	73 8 5 85 9 6 15.1	123 18 0 87 13 0 12.8	639 86 7 87 12 1 12.7	516 68 7 87 12 1 12.7
	unisys1	88 29 20 64 21 15 35.8	83 29 16 65 23 12 35.2	101 48 16 61 29 10 38.8	58 12 5 77 16 7 22.7	54 22 10 63 26 12 37.2	82 45 14 58 32 10 41.8	466 185 81 64 25 11 36.3	466 185 81 64 25 11 36.3
Overall Totals		993 220 20 81 18 2 19.5	923 208 21 80 18 2 19.9	1160 309 16 78 21 1 21.9	606 64 5 90 9 1 10.2	617 137 20 80 18 3 20.3	1009 246 14 80 19 1 20.5		
Foreign System Totals		879 197 20 80 18 2 19.8	690 185 21 77 21 2 23.0	1022 282 16 77 21 1 22.6	606 64 5 90 9 1 10.2	617 137 20 80 18 3 20.3	762 211 14 77 21 1 22.8		

Legend:

#T	#F	#NA
%T	%F	%NA
% Un-Weighted Err		

Matrix tabulation of results for the Dec 94 *ATIS SLS Test Results* - Using Minimal/Maximal Scoring Criterion, for the Class (A+D) Subset.

Matrix columns present results for Test Data Subsets collected at several sites, and matrix rows present results for different systems.

Numbers printed at the top of the matrix *columns* indicate the number of evaluable utterances in the Test Data (sub)set from the corresponding site.

"Overall Totals" (column) present *results for the* entire Class (A+D) Subset for the system corresponding to that matrix row. "Foreign Coll. Site Totals" present results for "foreign site" data (i.e., excluding locally collected data) for the Class (A+D) Subset.

"Overall Totals" (row) present results accumulated over all systems corresponding to the Test Data (sub)set corresponding to that matrix column. "Foreign System Totals" present results accumulated over "foreign systems" (i.e., excluding results for the *system(s)* developed at the site responsible for collection of that Test Data subset.)

Table A17 ATIS SLS Results: Class (A+D) by Collection Site

COMPARISON MATRIX: McNEMAR'S TEST ON CORRECTLY ANSWERED UTTERANCES FOR THE TEST:
Dec 94 ATIS SLS Test Results - Using Minimal/Maximal Scoring Criterion
Class A+D

	att1	bbn2	cmu1	cmu2	mit_lcs2	mitre1	sri1	sri2	unisys1
att1		att1	same	same	att1	att1	att1	att1	att1
bbn2			cmu1	cmu2	mit_lcs2	bbn2	same	sri2	bbn2
cmu1				same	cmu1	cmu1	cmu1	cmu1	cmu1
cmu2					cmu2	cmu2	cmu2	cmu2	cmu2
mit_lcs2						mit_lcs2	same	same	mit_lcs2
mitre1							sri1	sri2	unisys1
sri1								same	sri1
sri2									sri2
unisys1									

Table A18 ATIS SLS Class (A+D) McNemar Significance Test Comparisons